semiconductor devices

N. R. SINHA RAY

SEMICONDUCTOR DEVICES

NILRATAN SINHA RAY, *M. Sc.* (*Cal.*), *Ph.D.* (*Edin*)
Professor, Department of Electronics and Electrical Communication Engineering, Indian Institute of Technology, Khargpur, India (Retd.)

THE NEW BOOK STALL
5/1 Ramanath Majumder Street
Calcutta 700 009 : *INDIA*

JANUARY, 1990

Acc No — 15141

Price : Rupees Seventyfive only

Published by
MOHENDRANATH PAUL
THE NEW BOOK STALL
5/1 Ramanath Majumder Street,
Calcutta-700 009

Printed by Kanakprabha Basu at K. P. Basu Printing Works,
11, Mahendra Goswami Lane, Calcutta 700 006

PREFACE

Over the years the need for a comprehensive book on Semiconductor Devices dealing with both Physics and Technology that can cater for the needs of students of various categories has been felt. The present book aims at fulfilling the requirement. It starts from the basic physics, stresses on the semiconductor principles and on the devices. It is meant as a textbook on device and technology for the undergraduate and first year Master level students. Students desirous of higher studies on semiconductor devices, technology or integrated circuits needs a groundwork which the present book can provide.

This book may be recommended for one semester course work with a little selection of articles done by the teacher. Study questions and reading references are included at the end of each chapter. Hints or answers to some of these are also given. According to the need of the class, the device physics or technology may be emphasised by appropriate selection of these articles.

A book of this kind owes much to the publications of previous authors. These are duly acknowledged at the end of each chapter. Particularly for diagrams and standard results the author acknowledges gratefully the help he received from the following books :

The Physics of Semiconductor Devices by S.M. Sze.

Topics on Solid State & Quantum Electronics of John Wiley & Sons Inc.

Laser Handbook (Vol. I) by F.T. Atecchi & F.O. Schulz-Dubois of North Holland Publishing Company and the publication organisations of the IEEE, USA.

IEE Stevenge ; Herts Bell Laboratories, Pergamon Press Ltd., The Institute of Physics, Bristol.

The author is grateful to Dr. N.K. Mishra of the Department of Physics and to his colleagues at the Department of Electronics and Electrical Communication Engineering particularly to Prof. N.B. Chakraborty, Prof. S.K. Lahiri, Prof. S. Rakshit, Prof. D.

Bhattacharya & Dr. Swapna Banerjee of the Indian Institute of Technology, Kharagpur for reading and correcting parts of the manuscript. He is thankful to Mr. B.R. Laha for his competent typing and to Mr. M. Bose for his patient draftsmanship in the preparation of the manuscript. The author expresses his sincere appreciation of the efforts of the publisher in publishing the book in a neat and attractive form.

1st. January
1990
Kharagpur :
India

N. R. Sinha Ray

To

my Mother

who inspired me in writing this book

CONTENTS

			Pages
CHAPTER :	1	CRYSTAL STRUCTURE	1— 9
	1.1	Crystals and their Classification ...	1
	1.2	Miller Indices ...	4
CHAPTER :	2	SEMICONDUCTOR PHYSICS ...	10—65
	2.1	Semiconductor Device Principles ...	10
	2.2	Electrical Conductivity ...	10
	2.3	Mathematical Formulation of Particle Motion (Band Theory) ...	13
	2.4	Number of Modes in each Allowed Band ...	19
	2.5	Velocity, Acceleration, Effective Mass	21
	2.6	Distinction between Metals, Insulators and Concept of Holes ...	26
	2.7	Effects of the three dimensional nature of Actual Crystals ...	28
	2.8	Many Electron Problem ...	32
	2.9	Fermi Function ...	32
	2.10	Application of Fermi Statistics ...	34
	2.11	Lattice Vibration and its Consequences	39
	2.12	Temperature Dependence ...	45
	2.13	Extrinsic Semiconductors ...	46
	2.14	Hall Voltage ...	49
	2.15	Position of Fermi Level in Extrinsic Semiconductor ...	52
	2.16	Diffusion and Current Equations ...	56
CHAPTER :	3	HOMO-JUNCTIONS	66—85
	3.1	P-N Junctions ...	66
	3.2	P-N Junction from Energy Band Approach ...	71
	3.3	Minority Injection and Boundary Concentrations ...	73
	3.4	Ohmic Contact ...	76
	3.5	Carrier Decay Processes ...	76
	3.6	Determination of Doping Profile ...	81

Pages

3.7 Determination of Built-in Voltage ... 82

CHAPTER : 4 P-N JUNCTION DIODE 86—132

4.1 Built-in Field ... 86
4.2 Generalised P-N Diode ... 88
4.3 Non-equilibrium Situation ... 90
4.4 Current Flow Equations ... 93
4.5 Transmission Line Analogy ... 94
4.6 Minority Carrier Distribution ... 96
4.7 D.C. Current Voltage Relation ... 99
4.8 Injection Efficiency ... 100
4.9 Solution of Continuity Equation ... 100
4.10 General Expression for Current ... 103
4.11 Transport Factor ... 104
4.12 Diffusion Limited Diode D.C. Case ... 105
4.13 A.C. Impedance of Diode ... 110
4.14 Stored and Reclaimable Charges ... 112
4.15 Switching Characteristics of P-N Junction Diode ... 119
4.16 Charge Storage and Fast Recovery Diodes ... 121
4.17 Avalanche Breakdown and Zener Breakdown ... 122
4.18 P-I-N Diode ... 125

CHAPTER : 5 METAL SEMICONDUCTOR JUNCTION 133—157

5.1 Elementary Theory ... 133
5.2 Improved Theory ... 139
5.3 Current Voltage Relation for Schottky Diode ... 142
5.4 Point Contact Diode ... 149
5.5 Space Charge Limited Diode ... 150
5.6 Varactor Diode ... 152

CHAPTER : 6 JUNCTION TRANSISTOR 158—176

6.1 Working Principle ... 158
6.2 Power Gain ... 160
6.3 Current Gain ... 162

Pages
6.4 D.C. Characteristics ... 165
6.5 Punch-through Voltage ... 168
6.6 B.J.T. Breakdown Characteristics ... 171

CHAPTER : 7 BIPOLAR TRANSISTOR A.C. OPERATION 177—204
7.1 Transistor A.C. Equivalent Circuit ... 177
7.2 Early Effect ... 179
7.3 Intrinsic α Cut-off Frequency ... 183
7.4 Extrinsic Cut-off Frequency ... 185
7.5 Common Base Extrinsic Parameters ... 187
7.6 Common Emitter Parameters ... 188
7.7 Equivalent Circuit ... 190
7.8 Low Frequency Power Gain ... 192
7.9 Unilateral Power Gain ... 194
7.10 High Frequency Considerations in CE Configuration ... 196
7.11 β Cut-off Frequency ... 196
7.12 Measurement of Transistor Parameters ... 198
7.13 High Frequency Equivalent Circuits ... 200
7.14 Maximum Frequency of Oscillation f_m ... 202

CHAPTER : 8 TRANSIENT RESPONSE OF BIPOLAR TRANSISTOR 205—224
8.1 Transient Times ... 205
8.2 Computational Models ... 207
8.3 Charge Control Equations ... 209
8.4 Measurement of the Switching Parameters ... 213
8.5 Physical Meanings of the Time Constants ... 216
8.6 Turn on Delay ... 218
8.7 Transistor Rise Time ... 219
8.8 Saturation Delay ... 221
8.9 Fall-time with Over-drive ... 222

Pages

CHAPTER : 9 JUNCTION FIELD EFFECT TRANSISTOR **225—242**
9.1 Junction FET d.c. Characteristics ... 228
9.2 Field Effect Diode ... 235
9.3 Small Signal Equivalent Circuit ... 236

CHAPTER : 10 SEMICONDUCTOR TECHNOLOGY **243—254**
10.1 Diffusion ... 243
10.2 Oxidation ... 246
10.3 Epitaxy 246
10.4 Photolithography ... 248
10.5 Metallisation ... 248
10.6 Monolithic Circuit Fabrication ... 250

CHAPTER : 11 MOS AND MOSFET **255—311**
11.1 MOS Structure ... 255
11.2 MOS Transistor ... 258
11.3 Output and Transfer Characteristics ... 262
11.4 Small Signal Theory ... 263
11.5 Semiconductor Charge and Potential in MOS ... 266
11.6 Composition of V_T ... 270
11.7 MOS Capacitance and its Equivalent Circuit ... 275
11.8 MOS Capacitance under Deep Depletion ... 279
11.9 Surface States and MOS-CV Characteristics ... 280
11.10 Effect of Back Gate Bias on V_T ... 282
11.11 Effect of Temperature on Current ... 283
11.12 Standard (111) Process of MOST ... 284
11.13 Load Driver Pairs ... 287
11.14 MOS Technology ... 300

CHAPTER : 12 TRANSIT TIME DEVICES **312—346**
12.1 Avalanche Region ... 314
12.2 Avalanche Region Equations ... 317
12.3 The Read Diode ... 318

Pages

12.4 Different Impatt Structures ... 327
12.5 Read Diode Efficiency ... 328
12.6 Large Signal Considerations ... 330
12.7 Trapatt Mode ... 334
12.8 Impatt Mounting and Testing ... 337
12.9 Noise ... 340
12.10 Baritt Diode ... 341

CHAPTER : 13 GUNN DEVICES 347—392

13.1 Velocity Field Characteristics ... 347
13.2 Current Voltage Characteristics ... 350
13.3 Time Evolution of Accmulation Charge ... 352
13.4 Imperfect Cathode ... 356
13.5 Butcher's Equal Area Rule ... 359
13.6 Domain Mode ... 363
13.7 Resonant Circuit Loading ... 370
13.8 Starting of LSA Mode ... 381
13.9 Mode Chart of Gunn Devices ... 382
13.10 TED Signal Processing ... 384

CHAPTER : 14 MESFET 393—409

14.1 Principle of Operation ... 395
14.2 Equivalent Circuit ... 401
14.3 Design Considerations ... 406

CHAPTER : 15 HETEROJUNCTION 410—427

15.1 Abrupt Anisotype Junction ... 410
15.2 Anisotype Junction Current Voltage Relation ... 413
15.3 Abrupt Isotype Junction ... 417
15.4 Graded Heterostructures ... 421
15.5 Minority Injcetion ... 424

CHAPTER : 16 428—456

16.1 Photoemissive Detectors ... 429
16.2 Photoconductive Detector ... 434
16.3 Photodiodes ... 438
16.4 Avalanche Photodiode ... 447
16.5 Heterojunction Diodes ... 453

Chapter 1

CRYSTAL STRUCTURE

Materials in solid state can be broadly classified into two groups—crystalline and non-crystalline or amorphous. While atoms or groups of atoms are arranged in regular three dimensional order in crystalline state, in non-crystalline or amorphous state, they are in random fashion without any long range order in the arrangement. Metals, alloys and many other inorganic and organic solids are crystalline. Glasses, plastics and many organic solids are some examples of amorphous materials. The knowledge of atomic arrangements in solid materials is essential for understanding the basic principles of solid state devices. Crystalline materials have a major role in fabrication of solid state devices and it is, therefore, felt necessary to describe the crystal structures in short in the beginning.

1.1 Crystals and their Classification

Atoms or groups of atoms in a crystal are arranged in a regular order in three dimensions and it is always possible to choose a unit parallelopiped which can be moved in certain directions by some fixed distances such that the atoms or groups of atoms find themselves in identical surroundings. Hence, we define three primitive translational vectors $\vec{a}$, $\vec{b}$, $\vec{c}$ coinciding with the three edges of the parallelopiped and define such a translation as

$$\vec{r}_p = p_1\vec{a} + p_2\vec{b} + p_3\vec{c} \quad \dots \quad (1.1)$$

where p_1, p_2 and p_3 are integers. The points defined by the vector $\vec{r}_p$ are called lattice points and the assembly of these points is called a space lattice. The vector lengths a, b, c are chosen such that one obtains the highest symmetrical parallelopiped called the primitive cell having lattice points only at the corners.

It is often convenient to find Unit cells which are larger than the primitive unit cell and suited more to match the details of the crystal structure. In such a unit cell, besides the corners, atoms can occur at the centres of certain faces or at the centre of the volume.

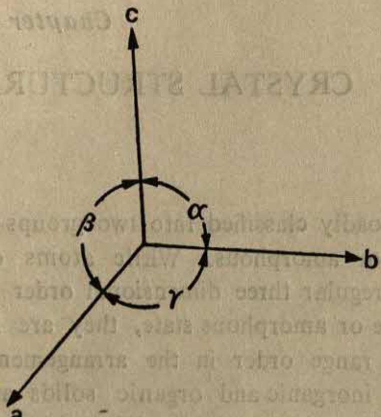


Fig. 1.1.1 shows the Unit cell vectors

If now new vectors $\vec{a}$, $\vec{b}$, $\vec{c}$ define the sides of the unit cell, the crystals can be classified into seven systems, depending on the relative dimensions of a, b, c and the angles included between them α, β, γ as shown in Fig. 1.1.1. They are : (1) triclinic, (2) monoclinic, (3) orthorhombic, (4) hexagonal, (5) rhombohedral, (6) tetragonal and (7) cubic. As shown in Fig. 1.1.2 (a) and (b), the different systems are described as :

(*i*) $c \leqslant a \leqslant b$ $\alpha \neq \beta \neq \gamma \neq 90°$ Triclinic

(*ii*) $c \leqslant a$ while b is arbitrary and $\alpha = \gamma = 90°$
$\beta \neq 90°$ Monoclinic

(*iii*) $c < a < b$ $\alpha = \beta = \gamma = 90°$ Orthorhombic type

(*iv*) $a = b \neq c$ arbitrary, $\alpha = \beta = 90°$; $\gamma = 120°$ Hexagonal

(*v*) $a = b = c$ $\alpha = \beta = \gamma \neq 90°$ Trigonal or Rhombohedral

(*vi*) $a = b \neq c$ $\alpha = \beta = \gamma = 90°$ Tetragonal

(*vii*) $a = b = c$ $\alpha = \beta = \gamma = 90°$ Cubic.

In this case, the translation of unit cells form the space lattice, and the space lattice is called (*i*) simple if the unit cell has got lattice points only at the corners, (*ii*) base centered if it has lattice points at the centres of a pair of opposite faces, (*iii*) face centered if it has lattice points at the centres of all the faces and (*iv*) body centered if it has lattice points at the centre of the body. This yields $4 \times 7 = 28$ possible classes. But not all of these are necessary to classify the crystals ; only 14 of them are adequate. This is because the rest can be grouped in either one of the above 14 categories with a redefinition of the unit cells. These fourteen Bravais lattices, as they are called after the name of their discoverer, are illustrated in Fig. 1.1.2 with some examples. These are : (1) Simple in triclinic ; (2) Simple and (3) base centred in monoclinic ; (4) Simple ; (5) base centred ; (6) face centred and (7) body centred in orthorhombic ;

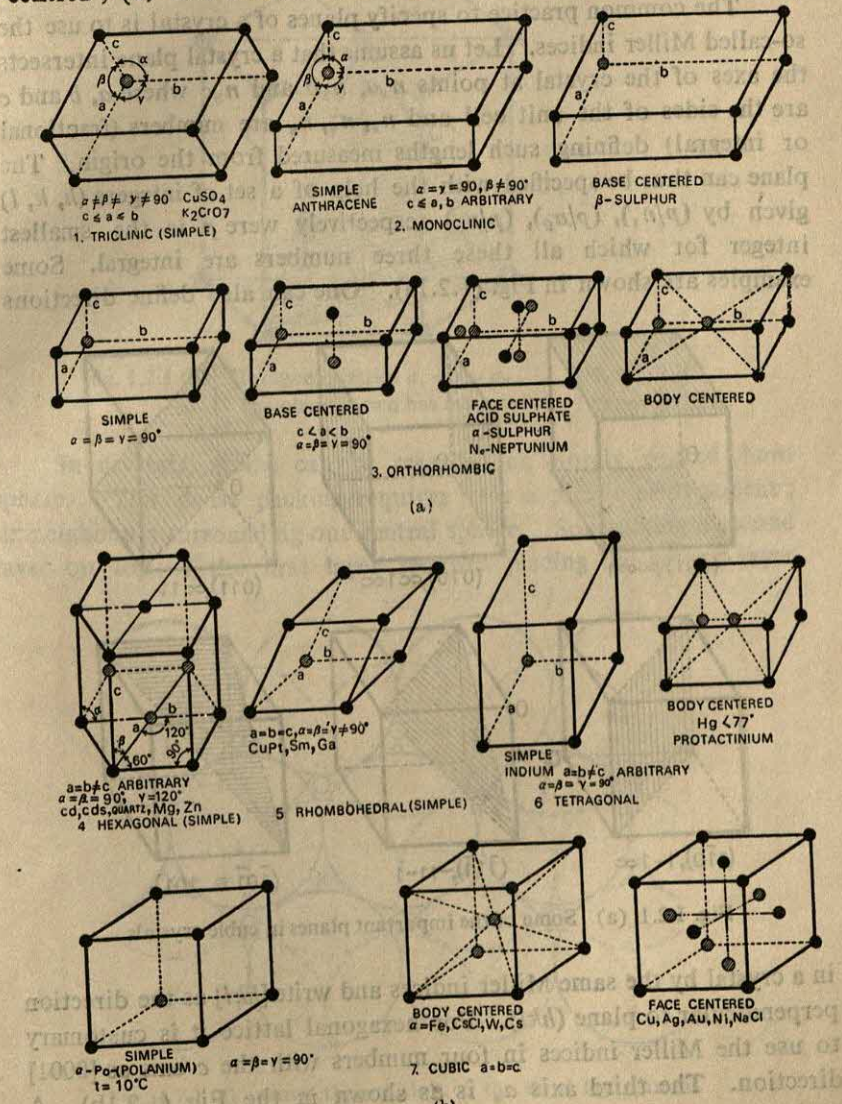
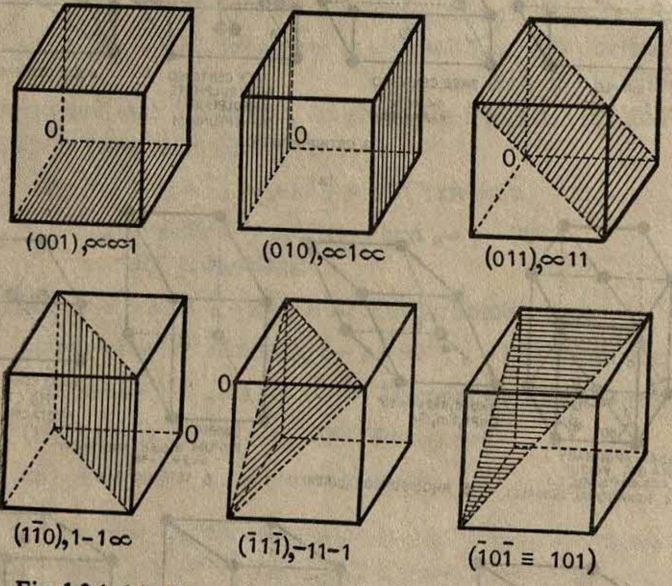


Fig. 1.1.2 (a) and (b) shows the fourteen bravais lattices

(8) Simple hexagonal ; (9) Simple rhombohedral ; (10) Simple and (11) body centred tetragonal ; (12) simple ; (13) body centred and (14) face centred cubic.

1.2 Miller Indices

The common practice to specify planes of a crystal is to use the so-called Miller indices. Let us assume that a crystal plane intersects the axes of the crystal at points n_1a, n_2b and n_3c where a, b and c are the sides of the unit cell and n_1, n_2, n_3 are numbers (fractional or integral) defining such lengths measured from the origin. The plane can thus be specified with the help of a set of integers (h, k, l) given by (p/n_1), (p/n_2), (p/n_3) respectively were p is the smallest integer for which all these three numbers are integral. Some examples are shown in Fig. (1.2.1a). One can also define directions in a crystal by the same Miller indices and write $[hkl]$ as the direction perpendicular to plane (hkl). For hexagonal lattice it is customary to use the Miller indices in four numbers with the c axis as [0001] direction. The third axis a_3 is as shown in the Fig. (1.2.1b). A typical plane is shown in Fig. (1.2.1b) in the hexagonal lattice in two



(001),∞∞1 (010),∞1∞ (011),∞11

$(1\bar{1}0)$,1−1∞ $(\bar{1}1\bar{1})$,−11−1 $(\bar{1}0\bar{1} \equiv 101)$

Fig. 1.2.1. (a) Some of the important planes in cubic crystals notations. The intercepts are $n_1 = 1$, $n_2 = 1$, $c = 1$ for (111) plane in the former, while $n_1 = 1$, $n_2 = 1$ and $n_3 = -\frac{1}{2}$ and $c = 1$ for $(11\bar{2}1)$ in the scheme of four axis notation.

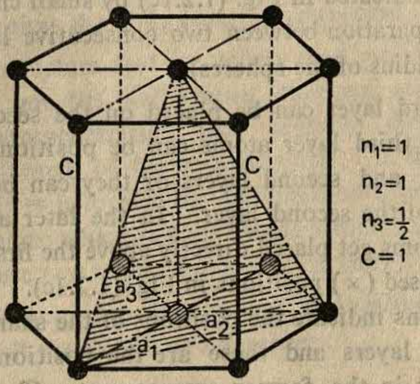


Fig. 1.2.1 (b) Hexagonal lattice $a_1 = a_2 = a_3$
The plane shown has intercepts at $\left.\begin{matrix} n_1 = 1, n_2 = 1 \\ n_3 = -\frac{1}{2}, c = 1 \end{matrix}\right\}$

In crystals, atoms can be regarded as closely packed hard spheres. This dense packing requires ; for a planar arrangement ; six neighbours surrounding one central sphere. So, packing a second layer on top of the first layer requires placing the second layer

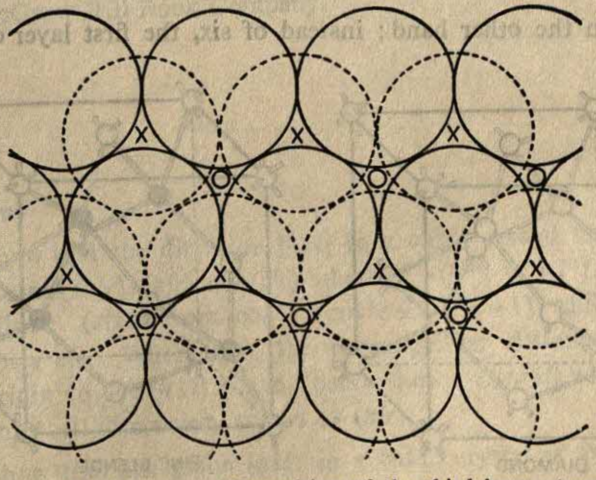


Fig. 1.2.1 (c) shows the position of the third layer atoms at crossed position for AB AB packing and at circled (o) position for ABC ABC packing

of atoms on the holes of the first layer. This leads to utilisation of only half of the number of holes (voids) of the first layer since parts of each alternate void of this layer remain as holes of the second layer as indicated in Fig. (1.2.1c) by small circles. It can be shown that the separation between two consecutive layers is R. $\sqrt{8/3}$ (where R is the radius of the spheres).

Next, a third layer can be placed on the second layer in two ways. Either the third layer atoms can be positioned on the holes of both the first and second layers or they can be placed on the alternative holes of the second layer. In the later arrangement, the third layer of atoms get placed directly above the first layer of atoms as shown by crossed (×) positions in Fig. (1.2.1c). In contrast, the circled (o) positions indicate the positions of the shared holes of the first and second layers and these are the positions of the atoms of the third layer in the former arrangement. If the third layer occupies the shared holes of the first two layers, then the fourth layer comes directly above the first layer of atoms. Thus, two basic types of atomic packing are obtained for close packing which are designated by the letters A, B and C as AB AB ··· and ABC ABC ··· . The former AB AB packing is called the hexagonal close packing while the latter [arrangement gives the cubic close packed structure as in face centered cubic (f.c.c.) crystals.

If, on the other hand; instead of six, the first layer of atoms

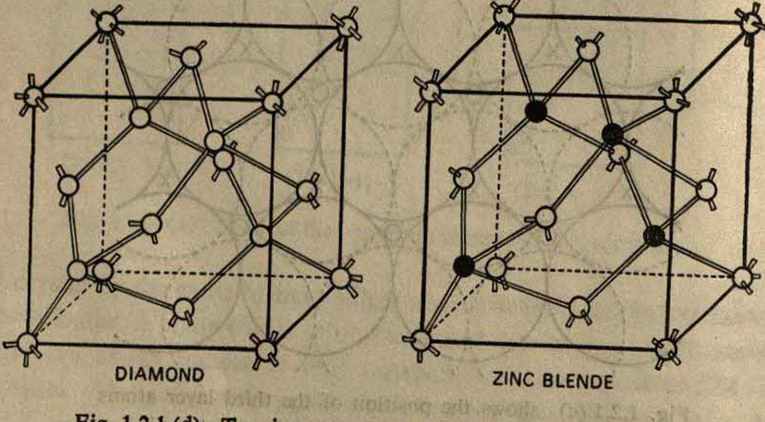


Fig. 1.2.1 (d) Two important unit cells, note the position of four typical gallium atom in zinc blende structure

has four nearest neighbours ; where, on the holes of the first layer the second layer of atoms are placed, the holes of the second layer occur directly above the first layer atoms. In this case the third layer of atoms get placed directly above the first layer of atoms. This arrangement gives the body centered tetragonal (b.c.t) structure.

The semiconductors that are widely used are Si, Ge, and GaAs. Si and Ge have got the well-known diamond structure. In this structure two interpenetrating face centered cubic lattices are displaced by one-fourth of the body diagonal. In the case of GaAs, one of the sublattices is formed by Gallium atoms, hence the structure has a different name which is Zinc blende structure, otherwise, it is like silicon. These two types of unit cell are shown in Fig. (1.2.1d).

For crystal bonding, defects and dislocations—the reader is required to consult the following references :

References

1. *Electronics of Solids* : W. R. Beam, 1965, McGraw Hill Inc.

2. *Solid State Physics* : A. K. Dekker, 1960, McMillan and Co. Ltd.

3. *Solid State Electronics* : W. Wang., Int. Student Edition, 1966, McGraw Hill Book Company.

Problems

1.1. Show that the direction [*hkl*] in a cubic crystal has direction cosines (h/r), (k/r) and (l/r) where $r^2 = (h^2 + k^2 + l^2)$.
Hint : (start from $\cos^2\alpha + \cos^2\beta + \cos^2\gamma = 1$) while $(d/n_1a) = \cos\alpha$ etc. Where d is the perpendicular distance form the origin to the (hkl) plane, hence show $r = (ap/d)$ where a is the unit cell dimension, so $\cos\alpha = (h/r)$.

1.2. Show that the plane (hkl) in a cubic crystal is at a distance (a/r) from the origin where a is the unit cell dimension. (Use p equal to unity in $r = (ap/d)$ of problem 1 above).

1.3. Find the distance between two nearest parallel planes for a cubic crystal corresponding to the planes (100), (110), (111) and (112).

[Ans. $a, a/\sqrt{2}, a/\sqrt{3}, a/\sqrt{6}$]

1.4. Calculate the distance between nearest neighbours in terms of unit cell dimension a in Si and GaAs. In which way are these neighbours of GaAs different compared to those of Si ?

[Ans. $(a.\sqrt{3}/4), a.\sqrt{3}/4$]

1.5. Find the number of atoms per cm^2 in the (100) and (111) planes in face centered cubic crystal where the unit cell dimension a is in angstroms.

[Ans. $2\times10^{16}/a^2$, $4\times10^{16}/\sqrt{3}.a^2$ since the lengths of the diagonals of faces is $\sqrt{2}.a$ while the (111) plane intersects these diagonals at mid-points, joining these mid-points gives another equilateral triangle each corner of which carries sixth of an atom within. The side length of the equilateral triangle is $a/\sqrt{2}$; hence the second answer follows.]

1.6. Find the angle between (100) and (111) planes ; (111) and $(\bar{1}21)$ plane.

Hint : Dot product of two unit vectors gives the cosine of the angle such as for the later pair

$$(i/\sqrt{3}+j/\sqrt{3}+k/\sqrt{3})\,(-i/\sqrt{3}+j/\sqrt{3}+k/\sqrt{3})=\tfrac{1}{3}.$$

[Ans. 54° 44′ and 70° 32′]

1.7. In the diamond structure see that each basic unit can be regarded as having four atoms situated at the corners of a tetrahedron formed by each of the cubes corner atoms with the three at the centres of faces. There are four atoms displaced quarter of body diagonal from the four typical corners. Assuming that these atoms touch the corner atoms, find the ratio of the volume occupied by the atoms to the total volume of the cube.

Hint : The side edges of the tetrahedron are $a/\sqrt{2}$. The distance of the displaced atoms from the corner atoms are $a.\sqrt{3}/4$. Find the atomic radius. Each of cubes corner atoms are shared by eight cubes, face centre atoms are each shared by two cubes and the four typical atoms are unshared ; hence find the number of atoms per cube equal to 8 and the required ratio as $\pi\sqrt{3}/16$.

1.8. In f.c.c. and hexagonal closed packed h.c.p. structures there are two types of holes in each structure. Between a corner and three face centre spheres packed in a tetrahedron there are tetragonal holes and between six spheres packed in an octahedron there are octahedral holes. Find the maximum radius of interstitial atoms that can be accommodated into these holes expressed as fractions of parent atoms radius in each case.

[Ans. f.c.c. tetragonal $\{(\sqrt{3/2})-1\}$ same as h.c.p. tetragonal and octahedral $(\sqrt{2}-1)$ same as h.c.p. octahedral.

Hexagonal closed packed structures hole locations are (1/3, 2/3, 1/4 and 1/3, 2/3, 3/4) for octahedral sites and for tetrahedral sites (0, 0, 3/8), (0, 0, 5/8), (7/3, 1/3, 1/8) and (2/3, 1/3, 7/8).]

1.9. How many tetragonal holes and octahedral holes are there corresponding to each unit cell including the shared ones in problem 8 above ?

Hint : One full octahedral hole is formed by six face centre spheres while there are twelve octahedral holes having one of the sides as body diagonal that are shared by four neighbouring unit cells ; hence in f.c.c. 4 octahedral. Similarly, 8 tetrahedral. In h.c.p. 4 tetrahedral and 2 octahedral.

Chapter 2

SEMICONDUCTOR PHYSICS

2.1 Semiconductor Device Principles

The first topic in any discussion of semiconductor devices should be the materials out of which these are fabricated. It is the physical properties of these materials as influenced by external agencies that are utilised in the working of a device. The most important physical property is the electrical conductivity either of the bulk material, of the surface of a material or a junction or junctions of materials which is profitably controlled by external agencies like electromagnetic radiation, electric field, magnetic field, pressure, temperature singly or in combination. In almost all devices, the electrical conductivity is controlled. The conductivity of different semi-conductor materials ranging from grey tin to diamond varies over many orders of magnitude; probably 10^{30} or so. Therefore it is worth enquiring as to what does determine this conductivity.

2.2 Electrical Conductivity

It is known that the charge carriers *viz.* electrons in atoms can have discrete energies. If they are excited, the electrons can jump from one energy level to another but cannot have energies in between these definite values. When such atoms arranged in a regular order make up the crystal, coupling between atomic orbits occur as shown in Fig. (2.2.1a) resulting in broadening of these discrete energy levels into bands as shown in Fig. (2.2.1b). Depending on the atomic spacing and the farthermost occupied orbital electron number in some crystals; the last occupied band becomes continuous. This is because, coupling causes splitting of an orbit into N numbers where, N is the number of atoms that have got coupled orbit and the next higher energy orbit has got more coupling, hence more splitting occurs. So the gap between the discrete orbits of the uncoupled atoms decreases on coupling. Thus coupling can give rise to continuous bands for higher energy orbits. If the last occupied band is continuous, the electrons move around from one atom to the other

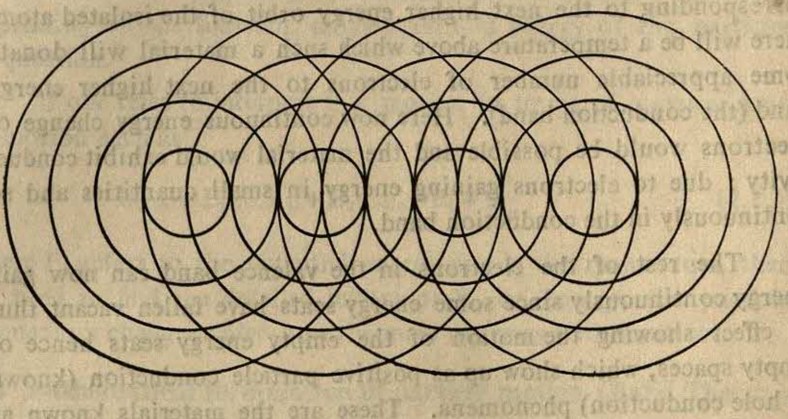


Fig. 2.2.1 (a) shows coupling of energy orbits

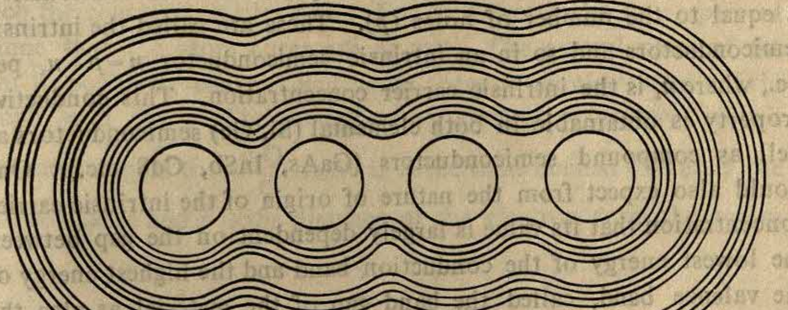


Fig. 2.2.1 (b) shows coupling of energy orbits resulting in energy bands

even at very low temperatures and such materials are the metals, while there can be materials in which at very low temperatures, the last occupied band (the valence band) is completely full of electrons. These are the so-called crystalline* insulators. Though in such a material electrons can move around from one atom to the other, unfortunately no net quantum of energy can be transferred by external electric field to the electrons of the crystal.

---

* The word crystalline is used here to exclude the amorphous materials from this discussion.

But depending on the closeness of the next higher energy band, corresponding to the next higher energy orbit of the isolated atom, there will be a temperature above which such a material will donate some appreciable number of electrons to the next higher energy band (the conduction band). Here now continuous energy change of electrons would be possible and the material would exhibit conductivity ; due to electrons gaining energy in small quantities and so continuously in the conduction band.

The rest of the electrons in the valence band can now gain energy continuously since some energy seats have fallen vacant thus in effect showing the motion of the empty energy seats hence of empty spaces, which show up as positive particle conduction (known as hole conduction) phenomena. These are the materials known as semiconductors. Therefore in an ideally pure semiconductor the number of electrons (n) per c.c. that take part in electrical conduction is equal to the number of holes (p). These are called the intrinsic semiconductors and so in an intrinsic semiconductor $n=p=n_i$ per c.c., where n_i is the intrinsic carrier concentration. This conductive property is obtainable in both elemental (Si, Ge) semiconductors as well as compound semiconductors (GaAs, InSb, CdS etc.). One would also expect from the nature of origin of the intrinsic carrier concentration that its value is largely dependent on the gap between the lowest energy of the conduction band and the highest energy of the valence band, called the band gap of the material as also the temperature so also pressure.

When we think of the details of the energy bands and the motion of particles therein for a three dimensional crystalline structure, the thing becomes quite involved. In that a particular group of particles moving with a particular momentum (hence direction) will find the nuclei on their wayside more often than in another direction (another momentum). Also particles having a particular energy will penetrate nearer the nuclei than one having less energy in the same location and moving in the same direction. When we apparently ignore these facts of detailed particle motion and try to interpret the effect of the application of external electric field on their somewhat freeness ; implies ignoring the internal microscopic forces which are sometimes helping the externally applied electric field of force (thus such particles may be expected to exhibit lighter particle mass) or are sometimes opposing (thus exhibiting a heavier particle mass) depending on the location, momentum and energy of the particles in question.

If one tries to interpret this motion in terms of classical laws of motion such as

$$E_{\text{total}} = E_{\text{potential}} + E_{\text{kinetic}} = E_b + \frac{p^2}{2m} \qquad \cdots \quad (2.2.1)$$

where E_b refers to the potential energy and p the momentum, then we need to find the detailed nature of E_b or how p, the particle momentum changes ; also how the mass can vary with location.

2.3 Mathematical formulation of particle motion (Band Theory)

The atoms in a crystalline lattice are arranged in a regular order. So let us examine the microscopic aspects of the space through which the charged particles flow. For that matter let us at present assume that the atom cores are at rest. The central mass of the atom is positively charged and hence there is an electrostatic field around each atom. The energy of an electron is negative in this field and is equal to $-\frac{q.Q}{r}$ where Q is the total nuclear charge, q the electron charge and r is the distance of the electron from the centre

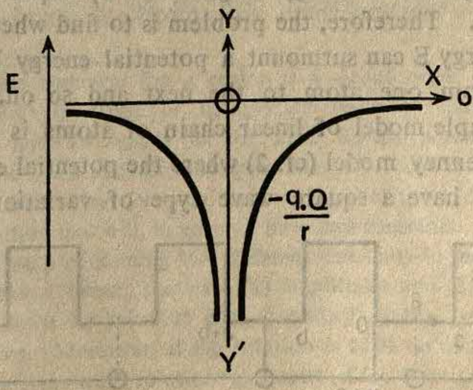


Fig. 2.3.1 (a) Potential energy around an isolated atom core

of the nucleus. This energy is zero at large distances from the atom. We are ignoring the effect of other electrons for simplicity. Therefore the microscopic variation of potential energy round an atom could be as shown in Fig. (2.3.1a) rotated round YY′ axis. The potential energy of an electron in an unidimensional crystal would therefore vary from point to point as shown in Fig. (2.3.1b) provided we ignore the effect of other electrons.

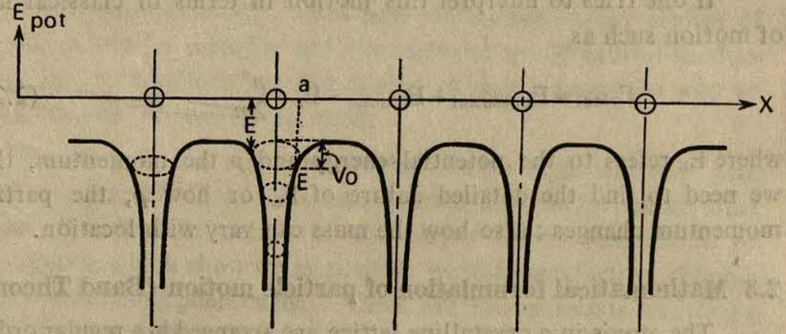


Fig. 2.3.1 (b) Potential energy of an electron in an unidimensional crystal

If the influence of other electrons are taken into account this potential energy variation would be more complex. Nevertheless since the atoms or atom groups of the same crystal are alike, there would exist periodic variation of potential in all directions while this period will, in general, be different in different directions. It will be noticed from Fig. (2.3.1b) that an orbital electron such as one at '*a*' would find a potential energy height $(V_0 - E)$ required for going to the next atom. Therefore, the problem is to find whether an electron having an energy E can surmount a potential energy barrier $(V_0 - E)$ and move from one atom to the next and so on. To treat this problem a simple model of linear chain of atoms is used, known as the Kronig Penney model (cf. 2) where the potential energy variation is assumed to have a square wave type of variation as shown in

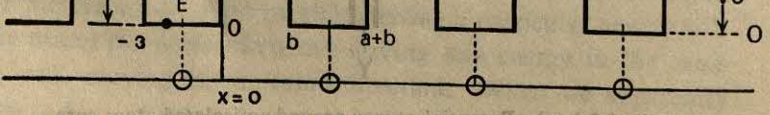


Fig. 2.3.2 Idealised model of periodic potential energy variation

Fig. (2.3.2). The electron having energy E is confined in one of these wells which has an energy height V_0. According to the classical mechanics it is not possible for the electron to Surmount the barrier since $E < V_0$. But it is known that according to wave mechanics the electron is regarded as a wave where the wave amplitude ψ describes by $\int_a^{a+d} \psi . \psi^* dx$, the probability of finding the electron between a to $a+d$, while $\int_{-\infty}^{+\infty} \psi . \psi^* dx = 1$ states that the electron is somewhere present. In this wave formalism the particle momentum is

$$p = \left(\frac{h}{2\pi}\right)\left(\frac{2\pi}{\lambda}\right) = \hbar k \qquad \cdots \quad (2.3.1)$$

where k is the phase factor of the corresponding plane wave $\exp [j(wt - kx)]$,

$$\text{that is } k = \frac{2\pi}{\lambda} \text{ and } h = 2\pi\hbar \qquad \cdots \quad (2.3.2)$$

Existence of the electron in the crystal implies that the electron has to stay as superposition of many plane waves, *i.e.*, a wave group where the group velocity will be the corresponding particle velocity. Hence one has to enquire whether at least one plane wave solution of the corresponding wave equations is possible in the two regions $0 < x < a$ and $-b < x < 0$, so that such plane waves can form the wave group. In this enquiry the wave amplitude continuity at boundaries and periodicity, *i.e.*, amplitude continuity after one period are invoked. From this consideration* one obtains that

---

*The wave equation for a region is a second order differential equation (cf. problem 2.3) in terms of the particle potential energy of the region. There are two types of regions (cf. Fig. 2.3.2) having potential energies V_0 in regions like $0 < x < a$ and zero in regions like $-b < x < 0$. Hence we have two differential equations one for each region. Each of these differential equations can have a solution that will, in general, have two constants. Hence we have two such solutions each containing two different constants in general. But we need one wave solution. Hence, the wave (1) amplitudes and (2) their slopes for the two solutions must be same at each boundary of the two types of regions mentioned above. Moreover, if the solution is to be an unattenuated wave since of necessity one plane wave solution must exist (if the solution is to exist) for the mode of description chosen, the wave (3) amplitude equality (*i.e.*, periodicity and there would have a solution *i.e.,* the electron can stay in the crystal as an unattenuated wave provided the equation (2.3.3) is satisfied

$$\frac{B^2+A^2}{2AB} \sinh Bb \sin Aa + \cosh Bb \cos Aa = \cos k(a+b) \quad \cdots \quad (2.3.3)$$

where $B^2 = \frac{2m}{\hbar^2}(V_0 - E)$ and $A^2 = \frac{2mE}{\hbar^2}$ ··· (2.3.4)

To easily visualise the implications of equation (2.3.3) in which $V_0 b$, which is a measure of the bondage of the electron to the particular nucleus is kept fixed while b is allowed to tend to zero giving

$$\frac{m(V_0 b)a}{\hbar^2} \; \frac{\sin Aa}{Aa} + \cos Aa = \cos ka \quad \cdots \quad (2.3.5)$$

Let $\frac{m(V_0 b)a}{\hbar^2} = M$ ··· (2.3.6)

and equation (2.3.5) is written as

$$M \frac{\sin Aa}{Aa} + \cos Aa = \cos ka \quad \cdots \quad (2.3.7)$$

Note from (2.3.4) that A^2 stands for the total energy of the electron.

The maximum and minimum limits of the right-hand side of equation (2.3.7) are ±1. Hence for a certain value of M one can draw the left-hand side versus (Aa) to obtain Fig. (2.3.3) which clearly shows that equation (2.3.7) can be satisfied provided (Aa) and hence particle energy has got values in certain ranges called allowed ranges and the electron and hence the wave will not stay in the crystal if the particle energy falls in the forbidden ranges.

The conclusions that we draw from Fig. (2.3.3) are

(1) The energy range consists of allowed and forbidden regions in succession.

(2) The width of an allowed range increases with increasing enrrgy or Aa ; this is because the amplitude of the first term of equation (2.3.7) decreases with increasing (Aa) on the average.

---

(4) continuity after one period of the periodic potential energy must hold between the solutions for the two types of regions. That is for solutions one period apart at a typical type of boundary the two types of wave, amplitudes equality and slope equality must hold. Hence, we see that there are four conditions to be satisfied by the two solutions, when we have four linear equations containing four constants (cf. problem 2.6).

(3) The width of particular allowed band decreases with increasing M, the bondage of the electron to an atoms core.

In one limiting cases M → 0, $\cos Aa = \cos ka$, hence all values of energy are allowed. From equation (2.3.4) with $A = k$ we have $E = \frac{\hbar^2 k^2}{2m}$ and all energy is kinetic corresponding to the case of free electron.

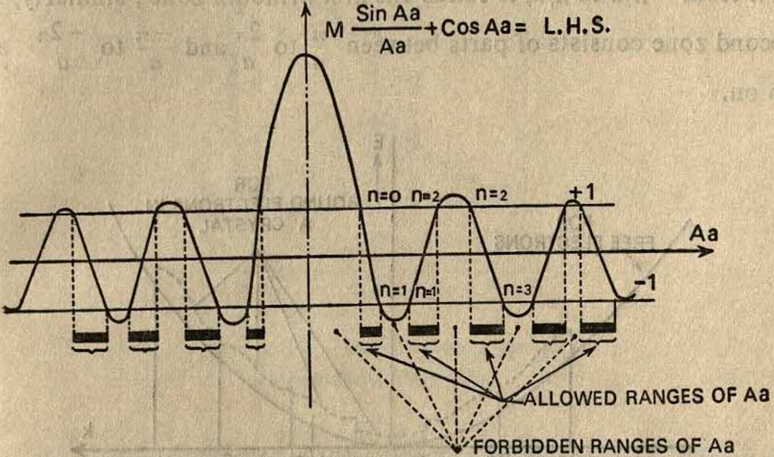


Fig. 2.3.3 A plot of $M\frac{\sin Aa}{Aa} + \cos Aa$ vs. Aa

In the other limiting case when M → ∞, $(\sin Aa = 0) = \sin(\pm n\pi)$ or $A = \pm\frac{n\pi}{a}$, and using (2.3.4) $E = n^2\pi^2\hbar^2/a^2 2m$, that is, the energies are discrete and correspond to the case of an electron bound in its orbit.

It is to be noted from Fig. (2.3.3) that the breaks in the allowed band occur where $\cos ka = \pm 1$, *i.e.*, at values of $k = \pm\frac{n\pi}{a}$, $n = 1, 2$, etc. Secondly, Fig. (2.3.3), is symmetrical about the vertical axis. Thirdly, that if k is replaced by $\left(k + \frac{2n\pi}{a}\right)$ in equation (2.3.7), the same values of energy are obtained as solutions of (2.3.7). Thus the phase factor has a periodic nature. This property is referred to as the property of k periodicity. For a free electron we know that

$E = \frac{\hbar^2 k^2}{2m}$, hence for a free electron the energy versus k curve is parabolic as shown in Fig. (2.3.4), *i.e.*, the energy increases with k hence with momentum. With this in view for a certain value of the binding strength M, E versus k curve can be drawn from equation (2.3.7), from which it is expected that E - k curve would be symmetrical about the E axis but is discontinuous as shown in Fig. (2.3.4) by dotted line curve. The extension of the (E - k) curve for ranges of k from $-\pi/a$ to π/a, is called the first Brillouin zone; similarly, the second zone consists of parts between $\frac{\pi}{a}$ to $\frac{2\pi}{a}$ and $\frac{-\pi}{a}$ to $\frac{-2\pi}{a}$, and so on.

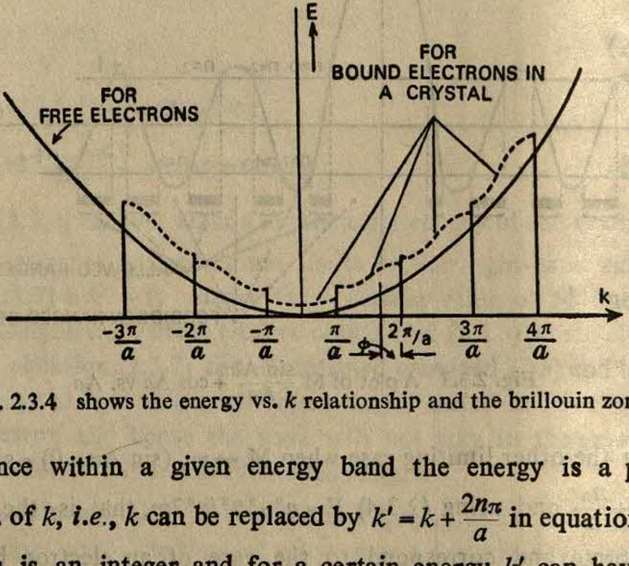


Fig. 2.3.4 shows the energy vs. k relationship and the brillouin zones

Since within a given energy band the energy is a periodic function of k, *i.e.*, k can be replaced by $k' = k + \frac{2n\pi}{a}$ in equation (2.3.7) where n is an integer and for a certain energy k' can have many values, therefore, it would be reasonable to represent the E versus k curves in one period, namely, between $-\pi/a$ to π/a. This representation would then be a representation in the reduced k space corresponding to $-\pi/a$ to π/a. For this we use the concept $\cos(2n\pi + \Phi) = \cos\left(\frac{\Phi}{a}\right).a$, so that the reduced phase space representation is as shown in Fig. (2.3.5). It should be noted that energy values at $\frac{2\pi}{a} \pm \frac{\Phi}{a}$ are epresented at $\pm\frac{\Phi}{a}$ in the k' scale. Secondly, because of the symmetry of the E - k curves, the slopes of the curves at $k'=0$ or at $\pm\pi/a$ are zero.

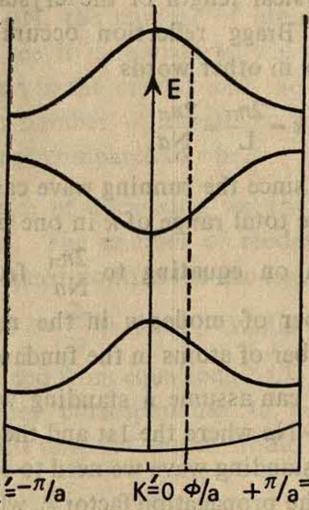


Fig. 2.3.5 Reduced phase space representation of E vs. k

In order to understand the behaviour of an electron within an energy band in the reduced wavefactor (k') space we need to know the following.

2.4 Number of modes in each allowed band

We have said that the particle electron is represented by superposition of plane waves. Therefore, it is natural to enquire how many of such plane waves, *i.e.*, modes of propagation are there.

It is well-known that it is the boundaries that determine the modes ; therefore, we need to look at the conditions at the boundaries. So far in our discussion we assumed a running plane wave, *i.e.*, infinite crystal length but in practice we shall always have finite crystal dimension. In addition, it is not known to us whether the ends of the crystal would cause reflection of the plane waves. Nevertheless it is known to us that there can only be two types of boundary conditions : (1) which will give a continuously running wave picture ; (2) the other which gives a standing wave picture.

If we assume that the former applies, the unattenuated plane wave propagation must give $\exp(ikx) = \exp ik(x + \mathrm{L})$ where $\mathrm{L} = \mathrm{N}a$ is some fundamental physical length of the crystal since we shall see later (cf. F.N.) that Bragg reflection occurs after such lengths comprising N atoms or in other words

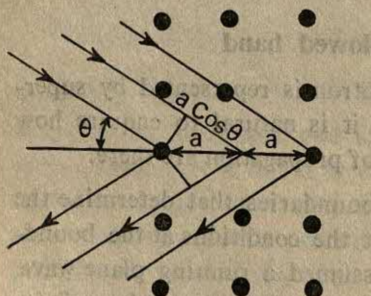
$$k\mathrm{L} = 2n\pi \text{ or } k = \frac{2n\pi}{\mathrm{L}} = \frac{2n\pi}{\mathrm{N}a} \qquad \cdots \quad (2.4.1)$$

where $n = \pm 1, \pm 2$ etc. since the running wave can be either a forward or a backward one. The total range of k in one period is $2\pi/a$, namely $-\pi/a$ to $+\pi/a$. which on equating to $\frac{2n\pi}{\mathrm{N}a}$ from equation (2.4.1), we have the total number of modes n in the running wave picture as N, equal to the number of atoms in the fundamental domain.

Alternatively, we can assume a standing wave picture within a fundamental length $\mathrm{L} = \mathrm{N}a$ where the 1st and the (N + 1)th atoms are the boundaries. For a standing wave we need to consider only positive or negative values of the propagation factor k, which ranges only over π/a. A standing wave solution will have nodes at the boundaries of the fundamental length $\mathrm{L} = \mathrm{N}a$ and will be given by the form

$$\sin kx = 0 = \sin n\pi = \sin k\mathrm{L} \text{ or } k = \frac{n\pi}{\mathrm{L}} \text{ where } n = 1, 2, 3, \text{etc.} \qquad (2.4.2)$$

F.N. : X-ray diffraction by a crystal points to a special behaviour of the wave phenomena. When a parallel beam of monochromatic X-rays of wavelength λ passes through a crystal, parts of it get reflected. The reflected rays from adjacent parallel planes can interfere constructively. This constructive interference occurs when the optical paths of reflected rays from the adjacent parallel planes differ by an integral multiple of λ. If the angle of incidence of the rays (cf. Fig. 2.4.1) is θ and the successive parallel planes are at distances 'a' apart, this condition, known as the Bragg reflection condition, can be written as $2a\cos\theta = \pm n\lambda$, where n is an integer. In terms of the phase factor, under this condition, the phase change between two successive parallel set of reflected rays is $ka = \pm n\pi$. We are not interested in light waves in particular but this wave property should be kept in mind when dealing with waves.



$$\mathrm{K}.\,a = \pm n\pi \left(= \frac{2\pi}{\lambda}\, a \text{ for } \theta = 0\right)$$

Fig. 2.4.1 Bragg reflection

Hence equating for the range of k, $\frac{\pi}{a}=\frac{n\pi}{L}=\frac{n\pi}{Na}$, we again have the number of modes $n=N$, the number of atoms in the fundamental length L. Thus we see it is immaterial whether a standing wave or a running wave persists in the crystal and the number of modes per band is equal to the number of atoms in the fundamental length $L=Na$ where N is large compared to one.

So, for the sake of discussion we shall henceforth use the running wave picture. The number of modes per band is too many and we can speak of quasi continuous modes ; since the total range of reduced wave factor $k=\frac{2\pi}{a}$ is finite. While each mode occupies a k range of $2\pi/L$ obtained from equation (2.4.1) by finding the distance between two k values corresponding to (n) and $(n+1)$. From equation (2.4.1) a small range of modes would occupy a small range of k given by

$$dn=\frac{L}{2\pi}\cdot dk \qquad \cdots \quad (2.4.3)$$

2.5 Velocity, acceleration and effective mass

Wave theory implies that the particle velocity v is given by the group velocity of the waves representing the particle. Therefore we have particle velocity

$$v=\frac{d\omega}{d\left(\frac{2\pi}{\lambda}\right)}=\frac{d(\hbar\omega)}{\hbar dk}=\frac{1}{\hbar}\frac{dE}{dk}, \qquad \cdots \quad (2.5.1)$$

since particle energy in wave formulation $E=hf=\frac{h}{2\pi}\cdot\omega=\hbar\omega$. Taking the E($k$) relationship in an allowed energy range, the velocity has been plotted as a function of k in the reduced zone in Fig. (2.5.1). From the plot the following two observations can be made.

(1) That starting from a momentum value $-\frac{\hbar\pi}{a}$, as the momentum increases the energy decreases while the velocity slowly becomes negative and maximum in magnitude at the inflexion point of the energy momentum curve. Then the velocity decreases to zero with increase of k, thereafter the velocity increase becomes positive

Acc No - 15141

maximum again at the E - k inflexion point and then decreases to zero at the top of the band at $k = \pi/a$,

(2) That even if the momentum values are finite, $\pm \pi/a$ at the band edges the particle velocity is zero, a fact not possible to be explained by Newtonian laws of mechanics. This momentum is called crystal momentum since such phenomena occurs when the particle is in the crystalline field.

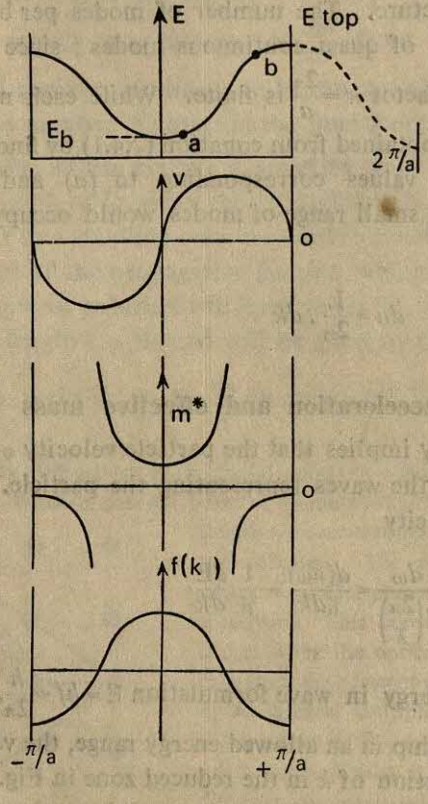


Fig. 2.5.1 Velocity, mass, acceleration obtained from reduced E vs. k relationship

This behaviour of the electron in the crystal is different from that we expect of a free electron and needs explanation which will be given at the end of this article.

We know that an electric field F applied to a particle would cause an increase of energy

$$dE = \frac{dE}{dk}dk = \pm eF\, v.dt = \pm \frac{eF}{\hbar} \cdot \frac{dE}{dk} \cdot dt \qquad \cdots \quad (2.5.2)$$

Note that, in equation (2.5.2), the plus or minus sign refers to the spatial direction though the energy is an increment derived from the field.

From equation (2.5.2), $\frac{dk}{dt} = \pm \frac{eF}{\hbar}$ ... (2.5.3)

It is apparent that k increases with time on application of electric field since the sign (plus or minus) in equation (2.5.3) does not refer to E $-$ k space.

From equation (2.5.1) the acceleration

$$a = \frac{dv}{dt} = \frac{1}{\hbar} \frac{d^2E}{dk^2} \cdot \frac{dk}{dt}$$

which, with (2.5.3), becomes

$$a = \pm \frac{eF}{\hbar^2} \frac{d^2E}{dk^2} = \pm \frac{eF}{m^*} \qquad \cdots \quad (2.5.4)$$

giving m^*, the effective mass as

$$m^* = \frac{\hbar^2}{d^2E/dk^2} \qquad \cdots \quad (2.5.5)$$

Thus the particle mass is given by the reciprocal of $\left(\frac{d^2E}{\hbar^2 dk^2}\right)$ and is obtainable from the E(k) relationship. In figure (2.5.1), for the particular E(k) curve chosen, the effective mass has been drawn and it is seen that between the bottom to the middle of the band the particle mass is positive, becomes infinitely heavy at the band centre and thereafter with increase of energy the mass is negative decreasing at the band top to a finite negative value.

The above mentioned apparently ambiguous results are, in essence, a consequence of the fact that ; ignoring the constraints on the particle motion imposed by the microscopic electrostatic forces of the nuclei of the crystal we were trying to understand the particle motion in the crystal, in terms of Newton's laws in respect of applied force ($\pm e$F). If the resultant microscopic force is opposing the applied force the mass would seem heavy because the acceleration will be lesser than accountable by the applied force alone and conversely. This, therefore, is a result of the electron not being free and being bound in the crystal to some extent. Therefore, it would be reasonable to find the extent of freeness of the particle as it moves in the energy band. Thus the effective freeness factor can be defined as

$$f(k) = \frac{m}{m^*} = \frac{m}{\hbar^2} \frac{d^2E}{dk^2} \qquad \cdots \quad (2.5.6)$$

implying that when the effective mass is large compared to the free electron mass m, the particle is behaving as a heavy particle. This is because the internal forces are largely dominating to dictate its motion and the particle is less free ; so the freeness factor is small. When $m = m^*$, the freeness factor is unity and the particle is behaving like a free electron.

Let us now try to understand the particle motion in the band when an external electric field is applied. Firstly we should note that on application of a force the momentum always gains since $\frac{dk}{dt}$ is always positive. This implies that the wave length λ shortens uniformly. Secondly that, the boundaries of the reduced k zone at $\pm \frac{\pi}{a}$; correspond to $\lambda = 2a$, which implies constructive interference occurring after reflection known as Bragg reflection, where the particle momentarily stopped ; indicating in the wave interpretation ; at this λ, standing waves are formed. An electron at b in Fig. (2.5.1) hence gradually is decelerated instead of being accelerated though gaining energy as it approaches the band edge with time. This accounts for velocity decrease or the negative mass necessary to express the requirement to be met. That is, if we want to disregard the strong internal forces of the crystal lattice and express the particle motion due to the applied force in terms of Newton's laws of motion the wave formalism guards such ignorance through an $E-k$ relation that sometimes requires a negative mass. For an electron on the lower half of the energy band the mass is positive. For motion from $k=0$ to values of k corresponding to energies like at 'a', momentum gains, velocity increases, so does the energy. The particle behaves like a partly free electron since it is in a crystal, the part freeness is taken care of by the increasing mass and its positive value.

After the particle has reached the zone boundary at reduced $k = \pi/a$, the momentum still increases but the energy falls [indicated by the dotted curve in Fig. (2.5.1), whose representation in the reduced k space occurs between $-\pi/a$ to $k=0$)]. The velocity increases negatively, so does the mass. This is a consequence of the fact that for continuous energy change of the electrons to occur ; it can only exhibit, to lower its energy since it is in the crystal. It has to remain as a wave group, of which some of the components wave length got shortened previously, (while the k range is finite), when some other components wavelegnth got increased. Only the role has got reversed. This is how the backward moving wave is exhibiting with it the particle, hence velocity increases from zero while remaining negative. Because we are trying to understand this by Newton's laws of motion, the consequence is of necessity, a negative mass giving a negative acceleration causing increase of negative velocity, in respect of the applied force. Thus the particle ultimately falling under the complete influence of an atom when the mass as seen by the applied force is infinite and velocity is a constant maximum and negative and the acceleration is zero. Because of the existing velocity the particle passes through this point thus the influence of the crystalline field decreases. The applied field feels its own influence on the particle as a decrease of the negative velocity towards zero. After this the velocity again becomes zero before it turns again to repeat the motion. Hence the mass required is positive and acceleration positive through zero to decrease the negative velocity to zero. Here the particle has stopped that is the wave group is reflected. The particle is well under the influence of the applied field, hence the mass is positive, so is the acceleration. It is to be noted that the mass is negative in the upper part of the energy band.

Thus we see that as if a steady electric field will cause an alternating particle motion in physical space hence will give alternating component of current in consonance with its oscillation in energy between E_{top} to E_{bottom} (see Fig. 2.5.1). For an energy band width of $2eV = E_{top} - E_{bottom}$ at a field strength of 100 v/cm, the physical distance should then be 0.02 cm. But the so-called mean free path is of the order of $10^{-5} - 10^{-6}$ cm, hence the electron will be deflected long before it could traverse a full cycle unless the field is $\simeq 10^5$—10^6 v/cm. However at these field strengths, there is a rapid increase of probability that the electron leaves this band and passes over to the next higher band. These aspects of collision with either lattice atoms or with imperfections were so far neglected since the ideal one dimensional fixed atom zone model was assumed.

2.6 Distinction between metals, insulators and concept of holes

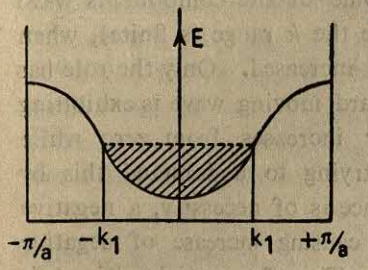


Fig. 2.6.1 shows states upto k_1 as filled

It would be instructive here to indicate how the above theory leads to a distinction between metals and crystalline insulators.

To understand this let us assume that an energy band has got filled up from the lowest energy upwards (cf. Fig. 2.6.1) to a value of $k = k_1$. We enquire : What is the number of free electrons with which such an energy band can be compared to ? This number is obviously

$$N_{\text{eff}} = 2 \sum_{-k_1}^{k_1} f(k).dn \qquad \cdots \quad (2.6.1)$$

where $f(k)$ is the freeness factor defined by equation (2.5.6) and dn is the number of modes in a range dk given by equation (2.4.3).

Hence $N_{\text{eff}} = \frac{L}{\pi} \int_{-k_1}^{k_1} f(k)\, dk$ since each mode can be occupied by utmost two electrons one of each spin (Pauli's principle). Substituting for $f(k)$ from equation (2.5.6)

$$N_{\text{eff}} = \frac{Lm}{\pi\hbar^2} \int_{-k_1}^{k} \frac{d^2E}{dk^2} \cdot dk = \frac{2mL}{\pi\hbar^2} \int_0^{k_1} d\left(\frac{dE}{dk}\right) = \frac{2Lm}{\pi\hbar^2} \left[\frac{dE}{dk}\right]_{k_1} \qquad \cdots \quad (2.6.2)$$

since $\frac{dE}{dk}$ at $k = 0$ is zero and $\frac{d^2E}{dk^2}$ is a symmetrical function. From equation (2.6.2) we can conclude that the effective number of electrons corresponding to a filled band is zero since at the top of the band $\frac{dE}{dk} = 0$. In other words a filled band does not contribute to conductivity. Also note that the effective number of free electrons reaches a maximum for a band filled upto the inflexion point of the $E - k$ curve.

Using the above information one can make a distinction between metals and crystalline insulators. At very low temperatures near absolute zero the electrons of the crystal will fill up the bands from the lowest upward until all the electrons are exhausted. There are materials for which the last occupied band is partly full and this shows conductivity. Such materials will be called metals. On the other hand there will be materials in which the topmost occupied band is completely full. Since a completely filled band is equivalent to zero number of free electrons, therefore such material will show no electrical conductivity at this temperature range. These will be called crystalline insulators in this temperature range. In this filled band the net charge movement is zero. That is if v_i is the velocity of the i-th electron the net charge movement is

$$-q\Sigma v_i = 0 \qquad \cdots \quad (2.6.3)$$

for all values of i in the band. As the temperature rises some electrons from the upper most filled band, called the valence band, will be excited and would be raised to the next higher allowed energy band called conduction band. There they will occupy the lower energy levels and would be able to gain energy continuously from the applied electric field and hence will show electronic conduction. While the vacant spaces, left in the valence band will be called holes. Their movement can be visualised as due to the net effect of the movement of the remaining electrons in the band which can be expressed from equation (2.6.3) as

$$-q\Sigma v_i = q.(v_j + v_k + v_l + v_m) \text{ etc.} \qquad \cdots \quad (2.6.4)$$

i stands for all except the few ($i \neq j, k, l, m$, etc.) that have left the band. Thus we see that the movement of the remaining electrons in the valence band is such that it is equivalent to movement of equivalent positive charges with the velocity of the vacant spaces in this band. This will also cause current conduction and we call this hole conduction in the valence band occurring due to the totality of flow of the remaining electrons in this band. It is obvious that the number of holes is equal to the number of electrons in such materials and these will be called intrinsic semiconductors. Thus the difference between a crystalline insulator and an intrinsic semiconductor is a matter of the number that are excited to make transition to next higher band. If the distance between the top level of the valence band and the bottom level of the conduction band is too large at normal temperatures the conducting holes and electrons would be too few, so the electrical conductivity exhibited by them would be too small to be counted and hence they will be called insulators. As for in diamond this energy gap, which we shall henceforth call bandgap, is 7 *e*V while in Ge, this value is 0.67 *e*V. Hence diamond is called an insulator while Ge a semiconductor.

It is to be noted that an electron in the upper part of the energy band has negative mass. Holes, therefore, at the top of the energy band should show positive mass. Not only that the hole mass would also be dependent on its location in the energy band.

2.7 Effects of the three dimensional nature of actual crystals

The one dimensional model that we have discussed so far is evidently an over-simplification, but this model can be extended to three dimension. There are other methods of approximation which also give us the same results except for a few more features that are easily understandable and these are as follows :

In an actual crystal the lattice spacing '*a*' will be different in the three orthogonal directions. Each of these physical directions will give us a *k* direction and a reduced *k* scale. Hence we need to regard *k* as a vector and we call it wave vector,

where $$k^2 = k^2_x + k^2_y + k^2_z \qquad \cdots \quad (2.7.1)$$

Not only this the boundaries of the Brillouin zone will then occur at $\vec{k} \, . \, \vec{r} = \pm n\pi$ (cf. Fig. 2.5.2). This means at values of

$$k_x = \pm \frac{\pi}{a}, \; k_y = \pm \frac{\pi}{b}, \; k_z = \pm \frac{\pi}{c}, \qquad \cdots \quad (2.7.2)$$

are the reduced zone boundaries, if *a*, *b* and *c* are orthogonal vectors.

This cubic region defined in k space forms the first Brillouin zone of a simple cubic lattice as well.

In view of the above, $k_x.b = 0 = k_y.c$ and so on. This leads us to define a set of vectors, A, B, C which specifies the unit cells of the k space. These are called the reciprocal lattice vectors expressed as,

$$\vec{A} = 2\pi \frac{\vec{b} \times \vec{c}}{\vec{a} \cdot \vec{b} \times \vec{c}}, \ \vec{B} = 2\pi \frac{\vec{c} \times \vec{a}}{\vec{b} \cdot \vec{c} \times \vec{a}}, \ \text{and } \vec{C} = 2\pi \frac{\vec{a} \times \vec{b}}{\vec{c} \cdot \vec{a} \times \vec{b}} \quad \ldots \quad (2.7.3)$$

The reciprocal lattice is therefore, a representation of the physical crystal in the phase space. The vectors $\vec{A}, \vec{B}, \vec{C}$ defined by equation (2.7.3) transform the physical crystal in the wave vector space or reciprocal space. Note that in the definitions of A, B, C only one length dimension remains in the denominator (cf. equation 2.7.2).

With these definitions the requirement corresponding to the equation (2.4.3) can be expressed for a three dimensional lattice as, the volume of k space to be ascribed for each mode is

$$dk = \frac{2\pi}{L_1} \cdot \frac{2\pi}{L_2} \cdot \frac{2\pi}{L_3} = \frac{(2\pi)^3}{V_f} \quad \ldots \quad (2.7.4)$$

where L_1, L_2, L_3 can either be taken as the three dimensions of the crystal or the three fundamental lengths at which the wave group suffers Bragg reflection.

The three dimensional nature of the actual crystal simultaneously puts a constraint on the allowed and forbidden energy bands as follows. An energy value E which is forbidden for a particular direction of propagation may be allowed for another direction of propagation. Thus if we consider all directions of propagation instead of a particular direction, the width of the forbidden band is reduced. This reduction can lead to an energy band forbidden for a particular direction of propagation to be nonexistent when all propagation directions are considered. We then say that band overlap has occurred.

With these introduction to three dimensional crystals we present in Fig. 2.7.1 the energy band diagram of three well-known semiconductors in two important directions in k space.

It would be seen that the semiconductors described above have not one but several conduction band minimum ; hence these are called multivalley semiconductors. The shapes of their constant energy surfaces are correspondingly shown in Fig. (2.7.2). For GaAs the constant energy surface is a sphere ; hence it will have one effective mass. For silicon the constant energy surfaces are six ellipsoids along the (100) axes and hence it will have three effective masses—one for longitudinal axis (m_l), two for transverse axes (m_t).

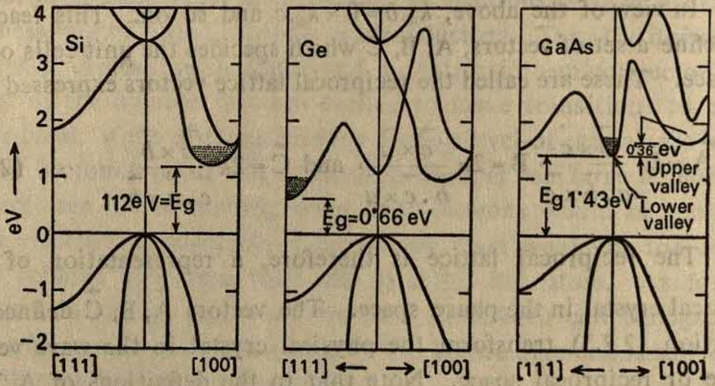


Fig. 2.7.1 Energy band diagram of Si, Ge and GeAs band gap values quoted in the figure refer to 300°K

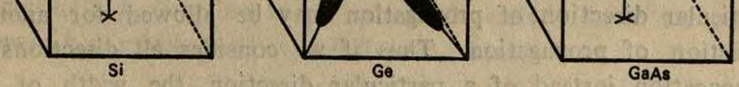


Fig. 2.7.2 Constant energy surfaces of Si, Ge and GeAs in reciprocal space. Brillouin zone boundaries are planes perpendicular to the major axes at the centres of the ellipsoids

Similarly, for Ge there are eight ellipsoids of revolution along the (111) axes. There will be also m_l and m_t. The Brillouin zone boundaries will be at the middle of the ellipsoids. The room temperature band gaps are shown on the diagrams (2.7.1). For these semiconductors the band gap decreases with rise of temperature. The conductivity effective mass may be calculated using the formula

$$\frac{3}{m^*}=\left(\frac{1}{m_1^*}+\frac{1}{m_2^*}+\frac{1}{m_3^*}\right)$$

In general, the energy is a function of k, *i.e.*, $E=f(k)$. It would be seen later that the holes and electrons near the band edges are the ones useful for conduction ; it is reasonable to express particle energy in terms of band edge energy $E_b=f(k_b)$, as

$$E=f(k)=E_b+\left.\frac{\delta E}{\delta k}\right|_{k=k_b}.(k-k_b)+\frac{1}{2}\cdot\left.\frac{\delta^2 E}{\delta k^2}\right|_{k=k_b}.(k-k_b)^2+\cdots$$

i.e., approximately

$$E=E_b\pm\hbar^2\left[\frac{(k-k_b)^2}{2m^*}\right] \quad \because \quad \left(\frac{\delta E}{\delta k}\right)_{k=k_b}=0 \qquad \cdots \quad (2.7.5)$$

The negative sign in the above expression refers to the lower part of the upper edge of the band and applies for holes where electron mass is negative. Equation (2.7.5) is an approximate equation, which tells that constant energy surfaces can be regarded as spheres in the k space. When the energy is measured with respect to the band edge and k is measured with respect to band edge $k=k_b$, the particle energy $E_1=E-E_b$ is related to the particle momentum by

$$\pm E_1=\pm\frac{\hbar^2 k_1^2}{2m^*} \qquad \cdots \quad (2.7.6)$$

which shows that E_1 can be regarded as totally Kinetic energy. The negative sign in equation (2.7.6) refers to holes which would be found below the E_b, *i.e.*, $E-E_b$ is negative.

Though, in general, constant energy surfaces are not spheres but for small deviations, it is customary to use this expression to frame a general picture for all semiconductors with suitable values of m^* determined from experiments.

Further, equation (2.7.5) indicates that the band edge energy can be regarded as the potential energy of a particle while at E_b and $\frac{\hbar^2(k-k_b)^2}{2m^*}$ is the Kinetic energy while the total particle energy E is sum of the potential and kinetic energies. Not only that the particles near the band edges can be regarded as free electrons with an effective mass, m^* given by equation (2.5.5), evaluated at the band edge. Thus, we note that the microscopic potential variation, within the semiconductor due to the presence of the nuclei, as if has disappeared giving us a microscopic description of the medium by equation (2.7.5), which implies that the particles in a band have a potential energy represented by the band edge energy E_b and a Kinetic energy represented by its momentum deviation from the band edge momentum.

2.8 Many Electron problem

So far we have considered one electron's behaviour in an energy band and it is not known to us how many such free particles are available in a band to account for conduction of current. Neither we know how they are distributed in energy to account for their Kinetic energies. To answer these questions, obviously, it is impossible to count the electrons one by one and ascribe them to different energies and we are compelled to take an account of all the particles in the total crystal volume by using statistical behaviour. The statistics that are applicable to electron or hole particles is Fermi-Dirac statistics This is because of Pauli's exclusion principle, that not more than two electrons can occupy each mode which ultimately points to the antisymmetric wave functions that have this property to represent these particles as opposed to Bose-Einstein statistics which takes account of Boson particles described by symmetrical wave functions.

2.9 Fermi function

Since we are not interested in rigorous procedures of derivation we shall be content with the engineer's way of understanding how the Pauli's exclusion principle enters to give us a function that describes the energy occupation probability for particles like electrons and holes.

Assume a closed system of large number of particles of at least two types, for example, electrons and the lattice which can exchange energy amongst themselves. Say, two such particles have energies E_{10} and E_{20} where their sum is $E=E_{10}+E_{20}$. Now say that these two particles have suffered an energy exchange and the new energies after collision are E_1 and E_2. Such a transition from one energy E_{10} to E_1 and E_{20} to E_2 will only occur provided the energy states E_1 and E_2 were empty as also E_{10} and E_{20} filled. This is where the Pauli's exclusion principle enters in the formalism. The probability of such an energy exchange will be determined by the probability of E_{10} and E_{20} being occupied and E_1 and E_2 are empty simultaneously. If the occupational probability of an energy state is represented by $f(E)$ then the probability of it being unoccupied is $[1-f(E)]$. The time rate of transitions of the above type can be written using a conversion constant C from probability to time rate as,

$$Cf(E_{10}).f(E_{20})\,[1-f(E_1)]\,[1-f(E_2)] \quad \cdots \quad (2.9.1)$$

Since the system is a closed one the total energy of the system can be regarded as constant, hence every microscopic process of transition of the sort, described above must accompany one of opposite transition. Similarly the time rate of reverse transitions will analogously be

$$Cf(E_1)\,f(E_2)\,[1-f(E_{10})]\,[1-f(E_{20})] \quad \cdots \quad (2.9.2)$$

and that $$E=E_1+E_2=E_{10}+E_{20} \quad \cdots \quad (2.9.3)$$

This allows us to write

$$E_1+\varepsilon=E_{10} \text{ and } E_2-\varepsilon=E_{20} \quad \cdots \quad (2.9.4)$$

where ε is an arbitrary energy deviation.

Equating now (2.9.1) and (2.9.2), dividing throughout by $f(E_1)\,f(E_2)\,f(E_{10})\,f(E_{20})$ and using (2.9.4) we have

$$\left[\frac{1}{f(E_1+\varepsilon)}-1\right]\left[\frac{1}{f(E_2-\varepsilon)}+1\right]=\left[\frac{1}{f(E_1)}-1\right]\left[\frac{1}{f(E_2)}-1\right] \quad \cdots \quad (2.9.5)$$

Equation (2.9.5) can be solved provided we substitute for the factor like

$$\left(\frac{1}{f(E)}-1\right)=Ae^{\beta E} \quad \cdots \quad (2.9.6)$$

which gives

$$f(E)=\frac{1}{1+Ae^{\beta E}} \quad \cdots \quad (2.9.7a)$$

The probability of occupation of very low energies are expected to be unity hence the constant A has to be very small and hence can be written as

$$A = (g/2)\exp(-\beta E_F) \qquad \cdots \quad (2.9.7b)$$

where g is called the degeneracy factor implying the number of particles populating a particular mode. Since two particles can populate one mode we have for electrons and holes

$$f(E) = \frac{1}{1 + e^{\beta(E-E_F)}} \qquad \cdots \quad (2.9.8)$$

Now particles having very high energies obviously will be too few when equn. (2.9.8) takes the form $f(E) = \frac{1}{A}e^{-\beta E}$. From gas laws we know $\beta = \frac{1}{kT}$ where k is Boltzmann constant and T the absolute temperature. Hence, equn. (2.9.8) takes the form

$$f(E) = \frac{1}{1 + e^{(E-E_F)/kT}} \qquad \cdots \quad (2.9.9)$$

To get a feel of the Fermi probability function described by equation (2.9.9) consider Fig (2.9.1) where $f(E)$ versus E has been plotted for two different temperatures.

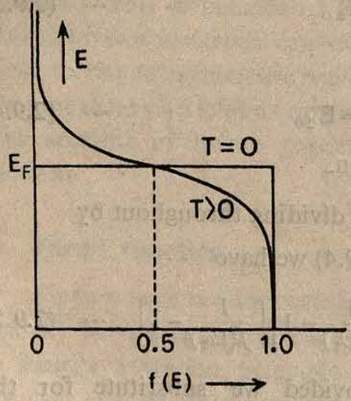


Fig. 2.9.1. Fermi probability function

For T = 0°K, at values of E less than E_F, $f(E) = 1$ and for values of E greater than E_F, $f(E) = 0$.

For small values of T,

at $E = E_F$, $f(E) = \frac{1}{2}$

at $E > E_F$ $f(E) = C\exp(-E/kT)$

at $E < E_F$ $f(E) =$ less than unity.

That is when the temperature rises from absolute zero, particles from below E_F rises above E_F and gets distributed in energy exponentially.

2.10 Application of Fermi Statistics

In order to apply the above probability function represented by equation (2.9.9) to account for the total number of electrons present in the solid we also need to know how densely packed the modes are as they are distributed in the energy scale. By this we mean how many modes are there ; in per unit energy range around the energy E, since the probability function $f(\mathrm{E})$ is a function of energy. If we know this density function D(E), then we can easily write the sum

$$2\Sigma \mathrm{D(E)}\, f(\mathrm{E})\, d(\mathrm{E}) = \mathrm{N} \qquad \cdots \quad (2.10.1)$$

for all electrons of all energies in the crystal or for those in the fundamental domain. The factor two comes in to take account of Paul 's principle and N is the total number.

The summation in equation (2.10.1) can be written with the notation for energy of the lowest edge of conduction band E_c as

$$\mathrm{N} = 2\sum_{-\infty}^{\mathrm{E}_1} \mathrm{D(E)} f(\mathrm{E})\, d\mathrm{E} + 2\sum_{\mathrm{E}_1}^{\mathrm{E}_2} \mathrm{D(E)} f(\mathrm{E})\, d\mathrm{E} + 2\sum_{\mathrm{E}_2}^{\mathrm{E}_3} \mathrm{D(E)} f(\mathrm{E})\, d\mathrm{E} \cdots\cdots$$
$$+ \sum_{\mathrm{E}_c}^{\infty} 2\mathrm{D(E)}\, \mathrm{F(E)}\, d\mathrm{E} = \mathrm{N}_1 + \mathrm{N}_2 + \mathrm{N}_3 \quad \cdots \quad (2.10.2)$$

where $-\infty$ and $+\infty$ indicate all energies are taken into account and N_1, N_2 etc. are the number of particles associated with each of the terms on the left-hand side of equation (2.10.2). Now considering that (1) in the forbidden gap $\mathrm{D(E)} = 0$, (2) that the lowest energy bands will be completely filled up, thus not contributing to conductivity and (3) Fermi function is a sharply decreasing function above $\mathrm{E_F}$ while the conduction electrons have been excited from the filled valence band, hence they are expected in the conduction band due to the Fermi function's exponential tail, the corresponding terms on either side of equation (2.10.2) drop out leaving mainly the conduction band electrons when we have :

$$2\sum_{\mathrm{E}_c}^{\infty} \mathrm{D(E)}\, f(\mathrm{E})\, d\mathrm{E} = \mathrm{N}_n \qquad \cdots \quad (2.10.3)$$

where N_n is the total number of electrons in the conduction band in the volume of the crystal corresponding to the fundamental domain V_f.

To obtain the D(E) dE of equation (2.10.3) we take the help of equations (2.7.4) and (2.7.6). Equation (2.7.6) states that the constant energy surfaces are spheres in the k_1 space. Between two such spheres of energies E_1 and $\mathrm{E}_1 + d\mathrm{E}$, the k_1 space volume is $4\pi k_1^2 dk$.

Following equation (2.7.4) each mode occupies a volume of $\frac{(2\pi)^3}{V_f}$ in this space. Hence the number of modes in the annular space between k_1^2 to k_1+dk is $V_f \frac{4\pi k_1^2 dk}{(2\pi)^3}$. This can be expressed in energy, normalised by kT, with equation (2.7.5) since $dk_1 = dk$ and $dE_1 = dE$ as,

$$D(E)dE = V_f . \frac{2}{\sqrt{\pi}} \left(\frac{2\pi m^* kT}{h^2}\right)^{3/2} \left(\frac{E-E_b}{kT}\right)^{1/2} \left(\frac{d(E-E_b)}{kT}\right) \quad \cdots \quad (2.10.4)$$

Hence equation (2.10.3) can be written (with E_c the conduction band edge energy) as

$$\frac{V_f 2N_c}{\sqrt{\pi}} \int_{E_c}^{\infty} \left(\frac{E-E_c}{kT}\right)^{1/2} \frac{d(E-E_c)/kT}{\left[1+\exp\left(\frac{E-E_F}{kT}\right)\right]} = N_n \quad \cdots \quad (2\ 10.5)$$

where $N_c = 2\left(\frac{2\pi m^* kT}{h^2}\right)^{3/2} = 2{\cdot}5 \times 10^{19} \left(\frac{m^2}{m}\frac{T}{300}\right)^{3/2}$ per cu cm.

$$\cdots \quad (2.10.6)$$

From equation (2.10.5) using the number of electrons per unit volume $n = (N_n/V_f)$ we have

$$\frac{2}{\sqrt{\pi}} \int_{E_c}^{\infty} \left(\frac{E-E_c}{kT}\right)^{1/2} \frac{d(E-E_c)/kT}{1+\exp\left[(E-E_F)/kT\right]} = \frac{n}{N_c} \quad \cdots \quad (2.10.7)$$

Equation 2.10.7 can be written with $\eta = (E-E_c)/kT$ as

$$\frac{2}{\sqrt{\pi}} \int_{0}^{\infty} \frac{\sqrt{\eta}\, d\eta}{1+\exp\left[\eta-(E_F-E_c)/kT\right]} = \frac{n}{N_c} \quad \cdots \quad (2.10.8)$$

Thus we see that in order to keep the book keeping correct ; for all electrons in the crystal, the number of electrons in the conduction band per unit volume must be related to (E_F-E_c) by the relation (2.10.8). When the integral has been performed and limits put, $(E_F-E_c)/kT$ will remain as a function of $\frac{n}{N_c}$. Unfortunately, this integral cannot analytically be solved for the whole range of η, but for the two limiting cases : (i) when $(E_c-E_F) \gg kT$ and (ii) when $(E_F-E_c) \gg kT$.

When $E_c - E_F \gg kT$, E_c is above E_F and $(E_F - E_c)/kT$ is strongly negative, hence the integral in (2.10.8) takes the form

$$\frac{2}{\sqrt{\pi}} \exp(E_F - E_c)/kT \int_0^\infty \sqrt{\eta}\exp(-\eta)\,d\eta = \frac{n}{N_c} \quad \cdots \quad (2.10.9)$$

or
$$\exp[-(E_c - E_F)/kT] = \frac{n}{N_c} \quad \cdots \quad (2.10.10)$$

In this case E_F has been formulated such that E_F is below E_c and the electrons having energy above E_c therefore lie in the Boltzmann tail of the Fermi function, where N_c acts effectively as the basic density of the base level E_c and $n \ll N_c$. Fig. (2.10.1) shows the function D(E) and D(E) *f*(E) for this case. Equation (2.10.10) can be re-written as

$$E_F = E_c + kT \log(n/N_c) \quad \cdots \quad (2.10.11)$$

This is the case of non-degenerate semiconductors.

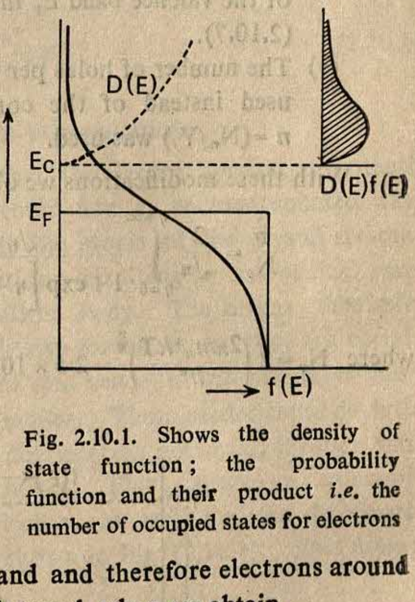


Fig. 2.10.1. Shows the density of state function; the probability function and their product *i.e.* the number of occupied states for electrons

The other limiting situation where $(E_F - E_c) \gg kT$ applies when E_F has entered the band and therefore electrons around E_F are the important ones in the integral, when we obtain

$$E_F = E_c + \frac{1}{2}\left(\frac{3}{8\pi}\right)^{2/3} \frac{h^2}{m^*} n^{2/3} \quad \cdots \quad (2.10.12)$$

This is the degenerate case or heavily populated situation when $n \gg N_c$ and corresponds to the situation in metals, check that for $(E_F - E_c)$, a few electron volts, $n \sim 10^{22}$/c.c.

Similarly to find the number of holes per unit volume available in the valence band that takes part in hole conduction, exactly similar book keeping can be made for all holes in the fundamental domain V_f paying due attention to the necessary modifications which are as follows :

(*i*) Since holes lie below the band edge the minus sign of equations (2.7.5) and (2.7.6) is relevant.

(*ii*) Since occupancy of holes implies non-occupancy of electrons, $[1-f(E)]$ has to be substituted for $f(E)$ in the equation like (2.10.1), in this case.

(*iii*) Also m_h^* is to be used for the effective mass of hole in the expression corresponding to (2.10.6) instead of m^* when it will be called N_v.

(*iv*) The integration would range from minus infinity to the top of the valence band E_v in the expression corresponding to (2.10.7).

(*v*) The number of holes per unit volume $p=(N_p/V_f)$ would be used instead of the corresponding electron case where $n=(N_n/V_f)$ was used.

With these modifications we obtain

$$\frac{p}{N_v}=\frac{2}{\sqrt{\pi}}\int_{\eta=0}^{\eta=\infty}\frac{\sqrt{\eta}\,d\eta}{1+\exp\left[\eta-\frac{E_v-E_F}{kT}\right]} \quad \cdots \quad (2.10.14)$$

where $N_v=2\left(\frac{2\pi m_h^* kT}{h^2}\right)^{\frac{3}{2}}=2{\cdot}5\times10^{19}\left(\frac{m_h^*}{m}\cdot\frac{T}{300}\right)^{\frac{3}{2}}$ per cm^3

$\cdots$ (2.10.15)

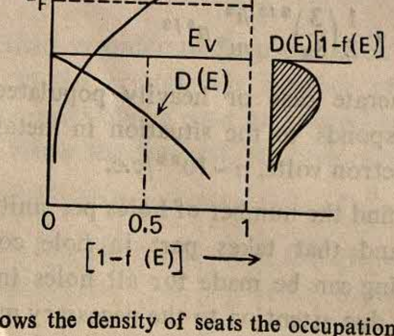


Fig. 2.10.2. Shows the density of seats the occupational probability and the number of occupied state for holes

and $$\eta = [(E_v - E)/kT] \quad \text{... (2.10.16)}$$

Hence likewise equation (2.10.14) gives for the corresponding non-degenerate hole concentration $p \leqslant N_v$ with the Fermi level being above E_v as,

$$E_F = E_v - kT \log (p/N_v) \quad \text{... (2.10.17)}$$

Fig. (2.10.2) depicts the corresponding density of state function and their occupancy when the Fermi level is above E_v. Similarly for the degenerate case when E_F is well below E_v and $p \geqslant N_v$ we have from (2.10.14)

$$E_F = E_v - \frac{1}{2}\left(\frac{3}{8\pi}\right)^{2/3} \frac{h^2}{m_h^*} p^{2/3} \quad \text{... (2.10.18)}$$

2.11 Lattice vibration and its consequences

So far we have assumed that the lattice (periodic) potential was stationary and the electron could flow as an unattenuated wave. Because of the temperature state the atoms of the crystal structure vibrate about their mean positions resulting in mechanical waves propagating through the crystalline body. The atoms themselves are bound to each other by cohesive forces and so they are coupled to each other by this spring force and hence the vibrational modes of mechanical waves are coupled modes. These modes cause deviation of the periodic potential, called deformation potential, through which the electron tries to flow as waves. This causes an interaction between the electron hence electron as wave and the deformation potential wave of the nature as shown in Fig. (2.11.1). This diagram is a crude oversimplification of the potential deviation produced by complex multimode mechanical waves, nevertheless it gives an insight as to how the energy exchange occurs. Electron in its motion at points like 'a' in Fig. 2.11.1 will be retarded by the field of the deformation potential wave while travelling through the crystal. Thus the electron energy would be decreased and the mechanical wave

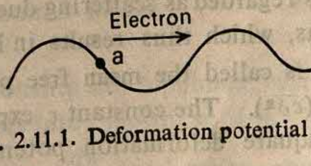


Fig. 2.11.1. Deformation potential wave

energy increased if the electron velocity is greater than the wave velocity and conversely.

The fixed lattice allowed unattenuated wave propagation representing the electron. The potential deviation would cause part reflection of this wave and part transmission. This will result in a decrease of the probability of finding the electron, becoming lesser and lesser as the wave goes forward. Hence the wave amplitude would decrease. The fractional decrease of the square of the amplitude, (*i.e.* the ratio of decrease of probability to initial probability) is expected to be proportional to the mean square of the amplitude (proportional to energy) of the deformation potential, hence

$$\frac{d|\psi|^2}{|\psi|^2} = -c\overline{\delta^2}dx \qquad \cdots \quad (2.11.1)$$

where c is the required proportionality constant. The solution of equation (2.11.1) is of the form

$$|\psi|^2 = |\psi_0|^2 \, exp\,(-c\overline{\delta^2}x) \qquad \cdots \quad (2.11.2)$$

The probability of finding the electron decreases exponentially with its distance of travel and is $(1/e)$ after a flight of length

$$l = (1/c\overline{\delta^2}) \qquad \cdots \quad (2.11.3)$$

The multimode mechanical wave motion can be regarded as motion of particles having particle energy $\hbar w$ corresponding to momentum $\hbar k$ where k is the phase vector and these are called phonon particles. The decreased probability of finding the electron at a further location is regarded as scattering due to collision between phonons and electrons, which thus results in hindrance to electron movement. In this l is called the mean free path which we see is equal to reciprocal of $(c\overline{\delta^2})$. The constant c expresses the relationship between mean square deformation potential δ to the electron energy loss. Any electron of energy E in the band will lose its kinetic energy on collision with the phonon particles. That is the multimode waves obtain energy from the electrons. This causes the individual atomic oscillators amplitude to increase. At a certain temperature of the lattice the oscillator stored energy on the average, is equal to kT and is proportional to the mean square displacement of the atoms from their positions of rest. This in turn is expected to be linearly related to the mean square deformation potential for small deviations at least. Hence we expect

$$l \propto 1/T \qquad \dots (2.11.4)$$

On the other hand if one analyses the longitudinal vibrational modes of (at least) a diatomic linear chain of atoms one finds for the frequency of vibrations in radians/sec

$$w = y[\,(1/M) \pm \{(1/M)^2 - (4 \sin^2 ka)/M_1M_2\}^{1/2}\,] \qquad \dots (2.11.5)$$

where M_1 and M_2 are the masses of the two types of atoms, $1/M = [\,(1/M_1) + (1/M_2)\,]$. y is the spring constant, 'a' the atom spacing and $k = \frac{2\pi}{\lambda}$ the phase factor.

From Equation (2.11.5) w has been plotted as a function of k in Fig. (2.11.2) between π/a to $-\pi/a$ since here also k is periodic. It is seen from this figure that there are two types of vibrational modes, the lower frequency modes are called acoustical mode while the higher frequency one is called optical mode. By the word acoustical, only lower frequency mechanical nature is meant but the frequency of this mode can be even in the microwave range, while optical modes extend to the optical frequency range. Note that there can be both transverse and longitudinal modes of these two types.

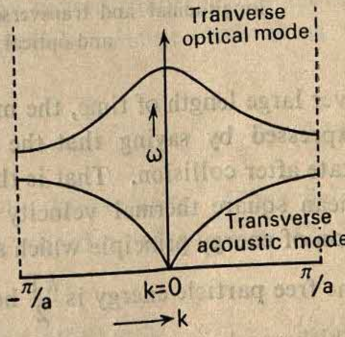


Fig. 2.11.2. Phonon mode dispersion curve of a diatomic linear lattice

The w, k relationship or phonon spectrum for a few practical semiconductors are shown in Fig. 2.11.3 and one can speak of phonon velocity $v = \frac{dw}{dk}$, acceleration, mass, number of modes. Obviously the number of modes in each branch will be equal to the number of atoms in the crystal, hence too many. While the mode range $\frac{2\pi}{a}$ is finite so quasi continuity of modes is assumed. Note that the quantum of energy exchanges at each collision of the involved phonon becomes $\hbar w \sim 10^{-14}$ ergs at the acoustic branch top end.

Since the electrons and phonons motions are not correlated, on collisions of such large numbers of them or of one particle seen

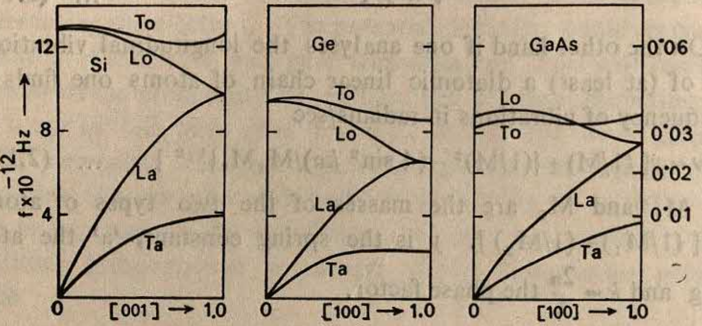


Fig. 2.11.3. Phonon spectra of three Typical Semiconductors in typical directions of normalised phase space.4 and T refer to longitudinal and transverse while *a* and *o* refer to acoustic and optical modes respectively

over large length of time, the motions remain uncorrelated. This is expressed by saying that the electron loses all memory of its past state after collision. That is the motion of electrons is random. The mean square thermal velocity can be estimated using the equipartition of energy principle which states that for each degree of freedom the free particle energy is $\frac{kT}{2}$ hence for three degrees of freedom $\frac{3kT}{2}$ which can be equated to the mean square K. E. of the electron to give the mean thermal velocity as $v_{th} = \sqrt{(3kT/m^*)} \sim 10^7$ cm/sec at room temperature range. Between two collisions free electrons travel a mean distance l given by equation (2.11.4). Hence for small deviations of periodic potential we can speak of a mean free time defined by the mean time that passes between two collisions of a particle and so this time

$$\tau = \frac{l}{v_{th}} \qquad \cdots\cdots \quad (2.11.6)$$

It is to be noted that we have tacitly assumed isotropic nature of τ (which is not strictly true). Anisotropic nature is taken care of in detailed transport theory, from which one obtains the usable effective mass (cf. prob. 2.8).

When an electric field is applied, the electrons gain energy from the applied field thus affecting the individual velocity components between collisions. So long the change is small compared to v_{th} the r.m.s. velocity can still be regarded as (v_{th}) remaining unaffected since on each collision the velocity is reduced to zero. Under this condition the mean free time τ can also be regarded as constant.

The effect of the modification of the individual velocity components and the collisions, results in an enmass drift of carriers moving with a drift velocity v_d. The kinetic energy that the mobile charged particles gain between collisions is given to the lattice by collision with acoustic mode phonons. The electrons are thus restored to the temperature of the lattice. Since the number of collisions is $(1/\tau)$; the rate of change of momentum per second is $(m^* v_d)/\tau$ due to collisions. This is the part of the applied force that is spent up to overcome the collision caused hindrance to motion. The rest goes to accelerate the particle. Hence with E as the electric field we can write

$$-q\mathrm{E} = m^* \frac{dv_d}{dt} + m^* \frac{v_d}{\tau} \qquad \cdots \quad (2.11.7)$$

to describe the effect of the field on the drift velocity. If the field E was abruptly applied at time $t = 0$, solution of equation (2.11.7) gives the time variation of v_d as

$$v_d = -\left(\frac{q\tau\mathrm{E}}{m^*}\right)[1 - \exp(-t/\tau)] \qquad \cdots \quad (2.11.8)$$

That is the drift velocity settles down to a value

$$v_{do} = -\left(\frac{q\tau}{m^*}\right)\mathrm{E} \qquad \cdots \quad (2.11.9)$$

with a time constant τ.

On the other hand if the field is withdrawn abruptly the velocity changes as

$$v_d = v_{do} \exp(-t/\tau) \qquad \cdots \quad (2.11.10)$$

i.e., the drift velocity falls to zero with the same time constant, so the particle velocities relax to their thermal equilibrium state with the time constant τ. Hence we speak of τ as the velocity relaxation time. The ratio of the steady state drift velocity to the electric field from equation (2.11.9) is called the mobility μ which from equation (2.11.9) is given by

$$\mu = \frac{q\tau}{m^*} \quad \cdots \quad (2.11.11)$$

and is a constant so long the applied fields are small such that the velocity deviation of individual flights are much smaller compared to v_{th}. To get a feel of the order of magnitude of times and distances involved, using $\mu = 1000$ cm²/v sec. and $m^* = 0.2\,m$ we have $\tau \sim 10^{-13}$ secs and the mean free path $l = v_{th}, \tau \sim 100$ A°. The velocity field characteristics of silicon, germanium and gallium arsenide are shown in Fig. 2.11.4.

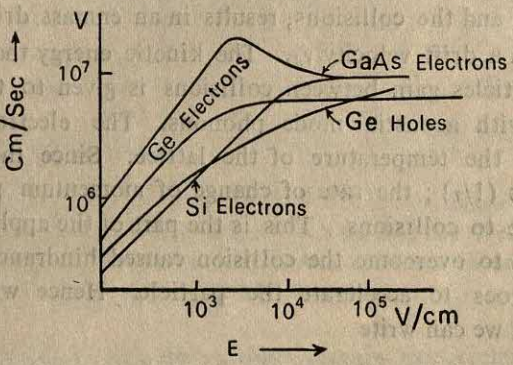


Fig. 2.11.4. Velocity field characteristics of Si & Ge

When of course the field is large ($\sim 10^3\,V/$cm), the individual velocity increments in the direction of the applied force between two collisions increase and become so large that the average thermal velocity is modified. In this state the average energy of the electron is greater than before and we speak of the electron as hotter than the lattice. In this field range it is not proper to speak of a constant time of flight since the velocity has increased and more collisions are expected in the same time interval because the maximum phonon velocity can be of the order of 10^5 cm/sec while the electron velocity is approaching 10^7 cm/sec, therefore large number of phonon particles appear as if at rest, for large number of electron particles. Hence the free time of flight must decrease on the average. So we turn our attention to the more fundamental quantity, the mean free path, for the description of events for high field case.

The mean space travelled in time τ,

$$l=\frac{1}{2}a\,\overline{t^2}=\frac{qE}{2m^*}\tau^2$$

giving $$\tau=\sqrt{\left(\frac{2m^*l}{qE}\right)} \qquad \cdots (2.11.12)$$

hence the drift velocity $$v_d=\frac{l}{\tau}=\sqrt{\left(\frac{qEl}{2m^*}\right)} \qquad \cdots (2.11.13)$$

and the mobility $$\mu=\sqrt{\left(\frac{ql}{2m^*E}\right)} \qquad \cdots (2.11.14)$$

Thus we see that in this range of field the drift velocity varies as square root of electric field as shown in Fig (2.11.4).

In essence the conservation of energy as also conservation of momentum is to be satisfied in the collision between phonons and electrons. In the higher momentum range in the acoustic mode the number of modes are fewer causing less frequent collision which in turn makes the electrons still hot and on collision to obtain conservation of energy and momentum proper; the frequency of collision increases, *i.e.* the time of flight decreases. This goes on until the electron energies have increased such that it is capable of exciting optical phonons which can easily absorb large energies. A phonon particle in this range implies very much larger energies compared to one of the acoustic mode. In this state, whatever energy is gained between collisions, is again lost in exciting optical mode phonon and so the average particle velocity remains constant in this range. As the field still further increases the energy gained between collisions is so large that the optical phonons become inefficient to remove this energy, when, we speak of the phonon mode bypassing and the collisions that occur starts, ionising the atom cores and we speak of impact ionisation when electrons are raised to conduction band, and holes are liberated in the valence band.

2.12 Temperature dependence

From the above discussion on the effect of field on the mobility it will apear that increased frequency of collision is expected to decrease the mobility due to lattice scattering. Therefore if temperature rises we would expect the mobility to vary. We have obtained from equation (2.11.11)

$$\mu=\frac{q\tau}{m^*}=\frac{q}{m^*}\cdot\frac{l}{v_{\text{th}}} \quad \cdots \quad (2.12.1)$$

Substituting for l from equation 2.11.4 and for $v_{\text{th}}=\sqrt{\left(\frac{3kT}{m^*}\right)}$ we obtain

$$\mu \propto T^{-3/2} \quad \cdots \quad (2.12.2)$$

In materials like Ge and Si the temperature dependence of the mobility deviates considerably from the above temperature law. The reasons for the discrepancy being that intervalley and intravalley scattering are not considered here and that acoustic phonon scattering has been taken to be the important mechanism. Nevertheless in these materials equation (2.12.2) approximately holds for lattice scattering.

The characteristic of impurity scattering on the other hand is a $T^{3/2}$ power dependence of μ. At higher temperature electrons on the average move faster and hence are less deflected by the impurity ion cores. It turns out when there are two sources dictating the velocity relaxation time τ ; then $(1/\tau)=(1/\tau_1)+(1/\tau_2)$ where τ_1 and τ_2 are the individual process relaxation times.

Hence $$\frac{1}{\mu}=\frac{1}{\mu_1}+\frac{1}{\mu_2} \quad \cdots \quad (2.12.3)$$

The combined effect is a decrease of mobility with rise of temperature, the power law depending on the predominance of one cause over the other.

2.13 Extrinsic semiconductors

Extrinsic semiconductors are those in which in equilibrium situation, the number of electrons n per c.c. is different from the number of holes p, per c-c. Equating equations (2.10.11) and (2.10.17) and rearranging we have

$$np=N_c N_v \exp[-(E_c-E_v)/kT]$$
$$=N_c N_v \exp[-(E_g/kT)] \quad \cdots \quad (2.13.1)$$

Thus we see that under equilibrium situation the product np is constant, depends on the semiconductor type through Eg and effiective masses m_h^* and m_e^* and on temperature. At a certain temperature for a certain material $np = \text{const.}$ and is same as for an intrinsic semiconductor of the same material for which $n = p = n_i$. So for any semiconductor

$$np = n_i^2 \qquad \cdots \quad (2.13.2)$$

Hence we see that n_i is temperature dependent and material dependent through Eg.

Elemental semiconductors like Si and Ge belong to the fourth group of the periodic table. Their electrons are arranged in different shells and sub-shells as $\begin{matrix} 1s & 2s & 2p & 3s & 3p \\ 2 & 2 & 6 & 2 & 2 \end{matrix}$ and $\begin{matrix} 1s & 2s & 2p & 3s & 3p & 3d & 4s & 4p \\ 2 & 2 & 6 & 2 & 6 & 10 & 2 & 2 \end{matrix}$ with atomic numbers 14 and 32 respectively. The last p shells in each of them are wanting 4 electrons to saturate the sub-shell. Each atom shares one electron from four of its nearest neighbours and so these crystals are called co-valent crystals. GaAs is a compound semiconductor made up of Ga from 3rd and As from 5th group, hence this is called a III-V compound.

If in these semiconductors impurities are substituted for some atoms, say in Si or Ge, atoms of the 5th group such as As, Sb, P which have an extra electron (for example in As at. no. 33, in 4P there are 3 electrons in phosporous at. no. 15, in 3P there are three electrons) in the outermost sub-shell ; these extra electrons will find little attraction to their parent atoms. These electrons are at a considerable distance away from their parent atoms and they see as if each is embedded in a dielectric medium only attracted by its share of counterpart positive charge. Hence the maximum possible energy with which it is bound to the nucleus can be given by the equivalent hydrogen atom models ground state as

$$\{mq^4/32(\pi\varepsilon_0\varepsilon_r\hbar)^2\} \qquad \cdots \quad (2.13.3)$$

which is of the order of 100 mev. Hence around room temperature these electrons will get free of their parent atoms and move around in the crystal. Thus these impurities can donate one electron per atom. Hence elements of the 5th group are called donor atoms. Thus this sort of doping can increase the conductivity of the sample and this conductivity is called *n*-type and the semiconductor is called *n*-type semiconductor. Similarly, if atoms of the 3rd group such as In, B are used as substitutional impurity, they having one electron less, the outermost sub-shell (such as in In at. no. 49 in the 5*p* one less, in B at. no. 5 in 2*p* one less, in Ga at. no. 31 in 4*p* one less) will be wanting an electron to satisfy the bond. This want of an electron by one bond per impurity atom will and does, act as a hole ; which will be filled by a passing electron, thus making the hole move freely through the crystal. These types of impurities will be called acceptor type and the conductivity of the crystal will be due to positive mobile charge carriers and the sample the so-called *p*-type sample. Whether the sample is *p*-doped or *n*-doped under equilibrium situations, equation (2.13.2) holds hence in doped crystal one type of mobile carriers (the dopant type) will be increased in number hence these carriers will be called majority carriers while the other species will be decreased in number such that (2.13.2) holds and these carriers will be called minority carriers. For example if one atom of 5th group impurity per 10^6 parent atoms are introduced as substitutional impurity, then per c.c. there will be 10^{16} electrons and 10^4 holes in silicon since $n_i \sim 10^{10}$ per c.c. at room temperature in silicon while 10^{16} electrons and 10^{10} holes in Ge since $n_i \sim 10^{13}$ in Ge at room temperature. Thus enormous change of conductivity is possible by doping a semiconductor. The conductivity of a semiconductor will be determined by both the types of carriers and the current density will be given by

$$J = (n\mu_n + p\mu_p)\, qE \qquad \cdots \quad (2.13.4)$$

where μ_n and μ_p are the electron and hole mobilities respectively and E is the applied electric field. Comparing with $J = \sigma E$ where σ is the conductivity we have,

$$\sigma = q(n\mu_n + p\mu_p) \qquad \cdots \quad (2.13.5)$$

Normally μ_n is greater than μ_p by about a factor of 2 to 3 for silicon and Ge. Hence major part of the conductivity is determined by the majority carrier doping.

As for the position of the energy level of the dopant atom particles in the energy scale, since very small evergy is needed to ionise these atoms hence in the doped semiconductors these energy seats, one for each atom, are drawn in the forbidden gap, as shown in Fig. (2.13.1). For electron states they are below the conduction band by the ionisation energy while for holes they are above the valence band by its ionisation energy. Ionisation of the electron state implies that the donor state is empty, it has donated an electron in the conduction band, while ionisation of an acceptor state implies that the acceptor state has captured an electron thus liberating a hole in the valence band.

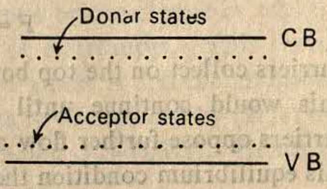


Fig. 2.13.1. Shows the valence band, the coduction band and the dopant states

2.14 Hall voltage

In this section we shall study the Hall effect which provides us with an experimental method for the determination of mobility, carrier concentration and the semiconductor type.

When a magnetic field B is applied to a *p*- or *n*-type semiconductor sample as shown in Fig. (2.14.1) in the *y* direction and

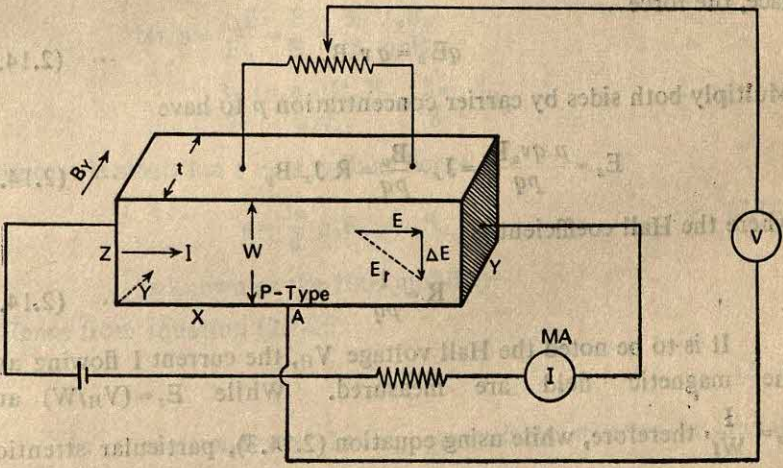


Fig. 2.14.1. Hall-effect measuring set-up

a current I amps $=(J_x.Wt)$ passes through the semiconductor perpendicular to the magnetic field in the *x* direction with carrier velocity v, the mobile particles get deflected due to a part of the Lorentz force

$$F = q(\vec{v} \times \vec{B}) \quad \cdots \quad (2.14.1)$$

Carriers collect on the top boundary perpendicular to the z direction. This would continue until the field produced by these collected carriers oppose further flow and an equilibrium is reached. Under this equilibrium condition there is a voltage developed on the top face with respect to the bottom one and this voltage is called the Hall voltage. This effect including the direction of the resultant field is called Hall effect. In the figure the change of field $\triangle E$ due to collected carriers is shown. The resultant field $E_r = E + \triangle E$ causes the carriers to travel a longer distance between the electrodes and this causes an increase in ohmic resistance of the sample. It turns out that fractional change in resistance hence the fractional conductivity change calculated on the basis of (J_r/E_x) is proportional to square of the magnetic field. This effect is known as magnetoresistance effect.

Turning our attention to the Hall effect, $\triangle E = E_z$ where E_z is the z directional field produced by the carriers collected at the top face, the force

$$qE_z = q\, v_x B_y \quad \cdots \quad (2.14.2)$$

Multiply both sides by carrier concentration p to have

$$E_z = \frac{p\, qv_x B_y}{pq} = J_x \cdot \frac{B_y}{pq} = R\, J_x\, B_y \quad \cdots \quad (2.14.3)$$

where the Hall coefficient is

$$R = \frac{1}{pq} \quad \cdots \quad (2.14.4)$$

It is to be noted the Hall voltage V_H, the current I flowing and the magnetic field are measured. While $E_z = (V_H/W)$ and $J_x = \frac{I}{Wt}$, therefore, while using equation (2.14.3), particular attention is to be paid to the measured quantities and the method of measurement. As for example to measure the Hall voltage V_H, one has to connect the measuring equipment between two points on opposite faces as shown in the diagram (2.14.1). The voltage measured is to be perpendicular to the current lines, hence the two voltage measuring contacts has to be accurately placed. Since it is difficult to do so, in practice one uses a resistance to form a bridge circuit in which the semiconductor bulk acts as a pair of resistances. This therefore gives the potential difference between point A and the midpoint of the semiconductor bulk. Hence W has to be replaced by (W/2) in the estimation of E_z from V_H but not in finding $J_x = (I/Wt)$. Hence from equation (2.14.3)

$$R = \frac{2V_H t}{I B_y} = \frac{1}{pq} \qquad \dots \quad (2.14.5)$$

In equation (2.14.3) the average drift velocity was used. But as a matter of fact the acceleration of the carriers due to the application of magnetic field is perpendicular to the direction of motion, hence due to collision little parts of circles become the nature of the mean free paths. Hence a more careful averaging should be done for the drift velocity. For semiconductors with spherical energy surfaces

$$R = \frac{3\pi}{8} \cdot \frac{1}{pq} \qquad \dots \quad (2.14.6)$$

Now using equation (2.14.3) the Hall angle defined by

$$\tan\theta = \frac{\Delta E}{E_x} = \frac{E_z}{E_x} = \frac{3\pi}{8}\,\frac{J_x B_y}{qpE_x}$$

$$= \frac{3\pi}{8}\,\frac{(\mu_p qp)E_x . B_y}{qpE_x} = \frac{3\pi}{8}\,\mu_p B_y \qquad \dots \quad (2.14.7)$$

Since θ is small $\tan\theta \simeq \theta$ radians hence

$$\theta = \frac{3\pi}{8}\,\mu_p B_y = \mu_h B_y \qquad \dots \quad (2.14.8)$$

μ_h is known as the Hall mobility

Hence from equation (2.14.6)

$$R = \frac{\mu_h}{\sigma} \qquad \dots \quad (2.14.8)$$

In a semiconductor having appreciable number of minority carriers R is given by

$$R = \frac{3\pi}{8}\,\frac{(p\mu_p{}^2 - n\mu_n{}^2)}{q(p\mu_p + n\mu_n)^2} \qquad \dots \quad (2.14.9)$$

and $$\mu_h = R\sigma = \frac{3\pi}{8}\,\frac{(p\mu_p{}^2 - n\mu_n{}^2)}{p\mu_p + n\mu_n} \qquad \dots \quad (2.14.10)$$

2.15 Position of Fermi level in Extrinsic Semiconductor

We are in a position to enquire where is the Fermi level located in the energy scale for intrinsic semiconductors and in doped semiconductors. For this we have the equations (2.10.11) and (2.10.17) which are applicable for both doped and undoped semiconductors. They are written again for ready reference

$$E_F = E_c + kT \log \frac{n}{N_c} \qquad \cdots \quad (2.15.1)$$

$$E_F = E_v - kT \log \frac{p}{N_v} \qquad \cdots \quad (2.15.2)$$

Adding the two equations we have

$$E_F = \frac{E_c + E_v}{2} + \frac{kT}{2} \log \frac{n}{p} \cdot \frac{N_v}{N_c}$$

and

$$= \frac{E_c + E_v}{2} + \frac{kT}{2} \log \frac{n}{p} \cdot \left(\frac{m_h}{m_e}\right)^{3/2} \qquad \cdots \quad (2.15.3)$$

Hence we see that the position of the Fermi level depends on the majority and minority carrier concentrations and their effective masses. But for intrinsic semiconductor $n = p = n_i$ whereas the logarithm of the mass ratio multiplied by $\frac{3}{4} kT$ would be small, hence for this case we can say E_F is about at the middle of the band gap.

As for doped semiconductors we need to know the carrier concentrations which are expected to depend on the doping concentration. But equations (2.15.1) or (2.15.2) do not explicitly show this. So we take equation (2.9.7b) in which we note that a donor level can be occupied by only one electron, hence the g factor for donor levels is unity. So the Fermi function that is applicable for donor level occupancy is

$$f(E) = \frac{1}{1 + \frac{1}{2} \exp\left[(E_d - E_F)/kT\right]} \qquad \cdots \quad (2.15.4)$$

Since donor levels have a discrete energy E_d as shown in Fig. (2.15.1). When E_F coincides with E_d, $f(E) = \frac{2}{3}$ hence the number of electrons in the conduction band obtained from N_d donor atoms is

$N_d\,[1-f(E)] = (N_d/3)$. The number of electrons for other positions of E_F as are obtained in the conduction band is

$$n = N_d\,[1-f(E)] = N_d\,[1/\{1+2\exp(E_F - E_d)/kT\}] \quad \cdots \quad (2.15.5)$$

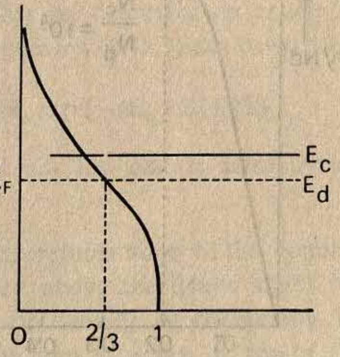


Fig. 2.15.1. The Fermi-function relative to the dopant energy state

By adding and subtracting E_c with the $(E_F - E_d)$ term in equation (2.15.5) and using equation (2.15.1) equation (2.15.5) can be written as

$$\frac{n}{N_d} = \frac{1}{\left[1 + 2\dfrac{n}{N_c}\exp\left(\dfrac{E_i}{kT}\right)\right]} \quad \cdots \quad (2.15.6)$$

where $E_i = (E_c - E_d)$ is the ionisation energy of the donor atoms. Equation (2.15.6) can be rearranged and written as

$$\frac{kT}{E_i} = \frac{1}{\log\left[\left(\dfrac{N_c}{2N_d}\right)\left\{\dfrac{(1-n/N_d)}{(n/N_d)^2}\right\}\right]} \quad \cdots \quad (2.15.7)$$

For a certain doping level N_d at a certain temperature, $\left(\dfrac{N_c}{N_d}\right)$ is known. Hence from equation (2.15.7) for a certain value of $\left(\dfrac{n}{N_d}\right)$, the ratio $\dfrac{kT}{E_i}$ can be calculated and plotted as shown in Fig. (2.15.2). From this figure we see that at values of $\dfrac{kT}{E_i} = 0.2$, almost all the donor atoms have got ionised. For Ge and Si the ionisation energies of different dopant atoms in, practice, range

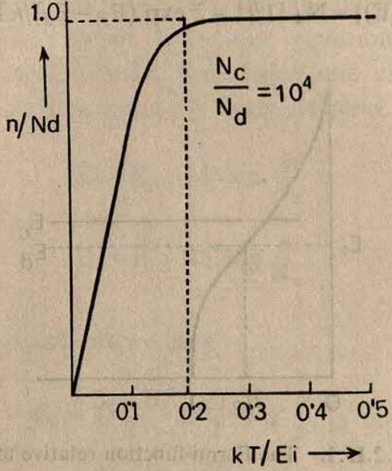


Fig. 2.15.2. Plot of (n/N_d) vs (kT/E_i)

between 10 mev to 150 mev. Hence we conclude that around the room temperature the number of conduction particles in the band is equal to the number of dopant atoms. These conclusions will not drastically change even if N_d is changed by a few orders of magnitude on either side because of the logarithm in the denominator of equation (2.15.7).

At lower temperatures the number of impurities ionised will be less and $\frac{n}{N_d} \leqslant 1$, in that case equation (2.15.7) can be rewritten neglecting $\left(\frac{n}{N_d}\right)$ in the term $\left(1 - \frac{n}{N_d}\right)$ as

$$n = \sqrt{N_d N_c}\, e^{-E_i/2kT} \quad \cdots \quad (2.15.8)$$

Equation (2.15.8) describes the number of electrons available in the conduction band of a doped sample at comparatively lower temperatures than room temperature.

Thus we see that the number of mobile carriers are low at low temperatures and tend to be equal to the dopant concentration around the room temperature. Whereas for all situations under equilibrium conditions, equations (2.15.1) and (2.15.2) hold. This is because the dopant concentration does not modify the band gap so long the doping is low and the sample is non-degenerate. (Atoms of the order of 1 in 10^4 or 10^6 are hardly noticed by the electron waves and their presence is taken care of by the modification in mobility as discussed in article 2.12). Continuing with our discussion, equation (2.15.1) can be written as

$$n = N_c \exp\left[-(E_c - E_F)/kT\right] \qquad \cdots \quad (2.15.9)$$

At low temperatures n is low and is given by equation (2.15.8) as also by (2.15.9).

At very low temperatures none of the donor levels are ionised hence E_F must be well above the donor states indicating that they are below E_F hence probability of occupation is unity. At high temperatures n is more than N_d since on top of all the donors being ionised some more has been excited from the valence band to the conduction band. Therefore at very high temperatures the material tends to behave like intrinsic again when the intrinsic concentration swamps those due to ionised donors, and so the Fermi level is near the mid gap. In between, when $n = N_d/3$, E_F is at E_d. Thus we see that Fermi level is temperature dependent in a doped semiconductor. At very low temperatures it is above E_d, with rise of temperature it slowly falls to the mid gap when the material behaves as if not doped. At low temperatures the concentration of electrons in the conduction band is given by equation (2.15.8), around room temperature $n = N_d$ while at high temperature the concentration would be given from equation (2.13.1) combined with (2.13.2) as

$$n = n_i = \sqrt{N_c N_v}\; e^{-Eg/2kT} \sim N_c \exp(-Eg/2kT) \quad \cdots \quad (2.15.10)$$

The log of concentration vers. $\left(\frac{1}{T}\right)$ is plotted in Fig. (2.15.3), using equations (2.15.10) and (2.15.8) for appropriate temperature ranges, which shows that there are two linear portions having slopes $-\tan^{-1}(Eg/2k)$ and $-\tan^{-1}(Ei/2k)$. We see that experimental plot of majority carriers versus $(1/T)$ curve can give us by extrapolation of appropriate portions, a means of determining band gap of a semiconductor as also the ionisation energy of a dopant atom.

From Fig. (2.15.3) we note that there is a temperature range from room temperature T_r to a temperature T_{cr} where the concentration remains constant $= N_d$ and the material behaves like a doped

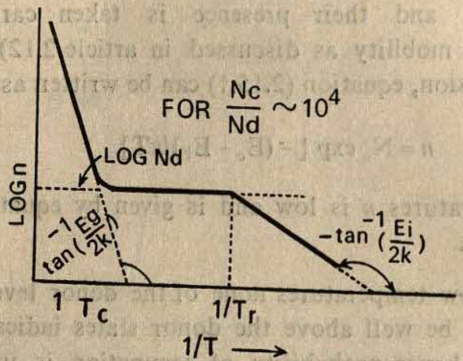


Fig. 2.15.3. Plot of the log of concentrations vers. (1/T)

semiconductor of *n*-type. It is this temperature range in which the semiconductor can be used reliably to behave as of a definite type. To estimate the value of this temperature range we have from equation (2.15.10)

$$N_d = N_c \exp(-Eg/2kT_{cr}) \quad \cdots \quad (2.15.11)$$

which on rearranging can be written as

$$\frac{T_{cr}}{T_{room}} = \frac{Eg}{2kT_r \log\left(\frac{N_c}{N_d}\right)} \quad \cdots \quad (2.15.12)$$

Using for $Eg = qV_g$

$$\frac{T_{cr}}{T_r} = \frac{V_g}{0.052 \log \frac{N_c}{N_d}}$$

With typical band gaps for Si and Ge one has for Ge, $T_{cr}/T_r \sim 1.2$ and for Si, $\frac{T_{cr}}{T_r} \sim 1.7$, hence we conclude $T_{cr} \simeq 60°C$ and $210°C$ above room temperature $T_r = 300°K$ for Ge and Si respectively.

2.16 Diffusion and current equations

Whenever there is a density gradient of carriers, there is a net flow of them from the high concentration to the low concentration region known as diffusion. This arises from the thermal velocities of the particles. For one dimensional movement we can characterise it as $\frac{m^* v_{\mathrm{th}}^2}{2} = \frac{k\mathrm{T}}{2}$. If we take a reference plane in the varying concentration region having, say, electron concentration n_0, the particles that cross this plane from either side, come from a distance of utmost a mean free path l. The concentrations at distances equal to the mean free path on either side of this reference plane can be written as $n_1 = n_0 + \left|\frac{\delta n}{\delta x}\right| \cdot l$, and $n_2 - n_0 - \left|\frac{\delta n}{\delta x}\right| \cdot l$. The net number that crosses the reference plane due to thermal movement per unit area per second is $\frac{n_1 - n_2}{2} = \left|\frac{\delta n}{\delta x}\right| \cdot l$. The current density is then $-q\left|\frac{\delta n}{\delta x}\right| \cdot l \cdot v_{\mathrm{th}}$ which with $l = v_{\mathrm{th}}\tau$ gives $-q\left|\frac{\delta n}{\delta x}\right| \tau v_{\mathrm{th}}^2$. Using $v_{\mathrm{th}}^2 = k\mathrm{T}/m^*$ and $\mu = q\tau/m^*$ we have for the current density by diffusion of carriers as $\mathrm{D}q\,\frac{\delta n}{\delta x}$ where $\mathrm{D} = (\mu k\mathrm{T}/q)$.

In a semiconductor without any voltage applied one would expect the same macropotential E_b to exist within the bulk material and the same concentration of carriers at each point hence from equations (2.15.1) or (2.15.2) we would expect E_F to be the same throughout the semiconductor bulk under equilibrium conditions. This is why we draw the conduction and the valency band edges as two parallel lines, indicated to represent constant macropotential along the length of the semiconductor. Even if there is a voltage applied across the semiconductor ends, and if no current flows then equilibrium has been attained and there is no further carrier movement hence E_F must have attained same value throughout the bulk ; since there is no net micro transitions necessary any further to satisfy equations (2.15.1) or (2.15.2), while E_F dictates the transitions. (For a formal proof the reader is referred to ref. 3). In such a case there can be macroscopic potential difference between different points and the representative for this macroscopic potential is E_b, hence representing $\mathrm{E}_b = -q\mathrm{V}$ and differentiating equation (2.15.1) we have

$$\frac{d\mathrm{E}_\mathrm{F}}{dx} = -q \cdot \frac{d\mathrm{V}}{dx} + \left(\frac{k\mathrm{T}}{n}\right) \cdot \frac{dn}{dx} \quad \cdots \quad (2.16.1)$$

Multiplying throughout by $n\mu_n$ and substituting $E = -\frac{dV}{dx}$ we can write equation (2.16.1) as

$$n\mu_n \frac{dE_F}{dx} = q\mu_n n\,E + (\mu_n kT/q).q.\frac{dn}{dx} \qquad \cdots \quad (2.16.2)$$

Representing $\frac{\mu_n kT}{q} = D_n$, we have from (2.16.2)

$$n\mu_n \frac{dE_F}{dx} = q\mu_n\, nE + D_n q.\frac{dn}{dx} \qquad \cdots \quad (2.16.3)$$

Under equilibrium condition no current flows and the L.H.S of equation (2.16.3) is zero and

$$E = -\frac{kT}{q}\ \frac{1}{n}.\frac{dn}{dx} \qquad \cdots \quad (2.16.4)$$

which states that if there is a field and there is equilibrium then the force due to the field on the carrier is exactly balanced by the force arising out of the gradient of their concentration. This force is known as the diffusion force and it is given by $\frac{kT}{n}.\frac{dn}{dx}$.

Under non-equilibrium conditions *i.e.*, when field is applied and the current flows, the current due to diffusion is independent of the field and it is entirely given by the existing concentration gradient, also the field current is *nq.n*E and is always so. Hence the total electron particle current is given by (2.16.3). Thus we have for electron current,

$$J_n = \mu_n qnE + D_n q\frac{dn}{dx} = n\mu_n.\frac{dE_{F_n}}{dx} \qquad \cdots \quad (2.16.5)$$

Similarly equation (2.15.2) gives

$$J_p = \mu_p\, q\, pE - D_p q.\frac{dp}{dx} = p\mu_p.\frac{dE_{F_p}}{dx} \qquad \cdots \quad (2.16.6)$$

Since there are two types of particles and the currents due to the individual types of carrier flow will in general, be different hence it is reasonable to use two Fermi levels. The above plausibility argument to establish the current under non-equilibrium situations is based on the idea that, under equilibrium conditions each component of current, diffusion and drift ; were equal and opposite and balancing each other to give zero resultant ; while the individual components may be large when we talk about unidirectional thermal current flow. Under non-equilibrium situations a slight unbalance of the two is expected to give appreciably large current. Hence we speak of small deviation from equilibrium and use the above current expressions. The smallness of the deviation again leads us to use the equilibrium Fermi level ideas to describe particle flow even under non-equilibrium situations. Bearing in mind a picture like, that Fermi level, as if acts as an elastic spring ; which under non-equilibrium forces gets deviated from equilibrium but relaxes quickly to restore the equilibrium, this phenomena occurring in quick succession gives the continuous current flow. Thus the Fermi level concept of equilibrium is extended to embrace non-equilibrium cases and under this formulation the non-equilibrium Fermi level is named by spelling it backward as 'Imref'.

References

1. *Physics of Semiconductor devices* : S. M. SZe, 1979.
2. *Solid State Physics* : A. J. Dekker 1960, p. 243, Macmillan & Co. Ltd.
3. *Electronic Semiconductors* : E. Spenke, 1958, p. 257, McGrawhill Book Co.
4. *Solid State Electronics* : S. Wang, Int. Student edition, McGraw Hill Book Company.

Problems

2.1. Starting from the one dimensional expression for a plane wave show that the corresponding particle wave function is $\psi = A_0 \exp[-j(Et - px)/\hbar]$. Argue to establish that sum total of such wave functions can represent a particle.

Hint : The particle energy in wave formalism is $E = \hbar w = K.E. + P.E.$ while particle momentum $p = \hbar k$ where $\hbar = h/2\pi$.

2.2. Find the one dimensional wave equation from the wave function obtained in problem 1 above.

Hint : Differentiate Ψ with respect to time to obtain $(\delta\Psi/\delta t) = -jE\Psi/\hbar$. Differentiate Ψ with respect to position twice to obtain $(\delta^2\Psi/\delta x^2) = -p^2\Psi/\hbar^2$ while K.E. $= p^2/2m$. Hence obtain $(\delta^2\Psi/\delta x^2) - 2m\,V\Psi/\hbar^2 + j(2m/\hbar)\frac{\delta\Psi}{\delta t} = 0$ where V is the potential energy of the particle.

2.3. Show that if the total particle energy is constant and the potential energy V is time independent the time independent form of the Schrodinger wave equation is

$$(\delta^2 X/\delta x^2) + (2m/\hbar^2)\,(E - V)\,X\,(x) = 0.$$

Hint : The product $E\Psi$ is linear in Ψ hence use variable separation method to obtain space and time variations separated by writing $\Psi = X(x).T(t)$. Equate the separating equation to time independent constant. Solve for time dependent part hence find $\Psi = X(x)\ \exp\ (-jEt/\hbar)$. Use $(\delta\Psi/\delta t) = -jE\Psi/\hbar$ to evaluate the constant as $-E$. Hence substitute and obtain the result.

2.4. Find the normalised wave function and the energies of a particle in one dimensional potential well of potential zero inside and infinite at the boundaries at $x = 0$ and $x = a$.

Hint : Start from the time independent form of the Schrodinger equation with $V = 0$. Find its solution which at the boundaries must be zero since the particle cannot have infinite energy hence find the condition of its existence in the well that gives the energies. Normalise the resultant $X(x)$ by the condition $\int_0^a X(x)\,X^*(x) = 1$; hence find the normalised $X(x)$.

[*Ans.* $E = (h^2 n^2 \pi^2/2ma^2)$ and $X(x) = \sqrt{\frac{2}{a}}\,(\sin n\pi x/a)$, $n = 1,2,3,$ etc.]

2.5. Find from the Schrodinger equation the orbital energy of the hydrogen atom and the atomic radius at ground state.

Hint : Use the three dimensional form of the Schrodinger equation $\Delta^2\Psi + \frac{2m}{\hbar^2}\ (E - V)\Psi = 0$ with the electron potential energy $V = -q^2/4\pi\,\varepsilon_0 r$. Assume θ and Φ symmetry, and obtain the equation for radial variation by separation of variables technique. Use a trial solution as $\psi(r) = A \exp(-r/r_0)$. Evaluate A by arguing that the probability of finding it somewhere between $r = 0$ to infinity is unity and use

$$\int_0^\infty \psi\, \psi^*\, 4\pi r^2\, dr = 1$$

that gives A, substitution of which gives $\psi(r) = 1/(\pi r_0^3 \exp 2r/r_0)^{1/2}$. Find the conditions, the trial solution should satisfy as $r_0 = \hbar^2\, \varepsilon_0/\pi\, e^2 m$ and $E = -(mq^4/8\, \varepsilon_0^2\, h^2)$ from the equation for radial variation with this solution $\psi(r)$. For further information on this problem the student should read reference 4, page 7.

2.6. Starting from the time independent Schrodinger equation for the one dimensional potential field shown in Fig. (2.3.2), deduce the equation (2.3.3) of the text.

Hint : Use the time independent form of Schrodinger equation given in problem (2.3) above with $X(x) = U(x) \exp jkx$ to take the effect of the periodicity of the lattice in modulating the plane wave into account. This is known as Bloch's theorem in the literature. This gives the equation $(d^2U/dx^2) + 2jk\; dU/dx - \left(k^2 - A^2 + \frac{2m\,V(x)}{\hbar^2}\right) U(x) = 0$ with $V = V_0$, $U(x) = U_1(x)$ for the region $-b < x < 0$ and $V = 0$, $U(x) = U_2(x)$ for the region $0 < x < a$.

A is defined in the text. For the two equations of these two regions write general solutions as $U_1(x) = C \exp j(A-k)x + D \exp[-j(A-k)x]$ and $U_2(x) = P \exp j(B-k)x + Q \exp[-j(B-k)x]$ and equate for the continuity conditions $U_1(0) = U_2(0)$ and $U'_1(0) = U'_2(0)$. Similarly, equate for the periodicity conditions $U_1(-b) = U_2(a)$ and $U'_1(-b) = U'_2(a)$. Thus four homogeneous equations are obtained for the four constants C, D, P and Q. These have non-trivial solution if the determinant formed by the coefficients is equal to zero. This gives equation (2.3.3).

2.7. The energy band gap of Ge, Si and GaAs varies with temperature according to the relation $E_g(T) = E_g(0) - \alpha T^2/(T + \beta)$. Where for Ge, $E_g(0) = 0.74$, $10^{-4}\alpha = 4.56$ and $\beta = 210$, for Si and

GaAs these constants are 1.16, 7.02, 1108 and 1.522, 5.8 and 300 respectively. Find the temperature coefficient of the band gaps around 300°K. Argue to establish that this should have been positive from the point of view of band theory.

[*Ans.* Ge : -3.78×10^{-4} *e*V/°K
Si : -2.67×10^{-4} *e*V/°K
GaAs : -4.35×10^{-4} *e*V/°K.]

2.8. Calculate the intrinsic carrier concentrations at 300°K of Ge, Si and GaAs using effective masses in relation to free electron masses for electrons and holes as for Ge $m_{le} = 1.6$, $m_{te} = 0.082$, $m_{lh} = 0.04$, $m_{hh} = 0.3$, for Si these are 0.97, 0.19, 0.16 and 0.5 while and for GaAs 0.068 $m_h = 0.5$ respectively.

Hint : Combine for electrons $m_{de} = (m_{le} \cdot m_{te}^2)^{1/3}$ and for holes $m_{dh} = (m_{lh}^{3/2} + m_{hh}^{3/2})^{2/3}$ to obtain the density of state effective masses taking account of parabolic nature of constant energy surface. Hence obtain

$$n_i = 4.9 \times 10^{15} \left(\frac{m_{de} m_{dh}}{m^2}\right)^{3/4} T^{3/2} \exp\left[-E_g/2kT\right].$$

For E_g use data from problem (2.7) above.

2.9. Find the ionisation energy of Zn atom when it replaces a Ga atom in GaAs crystal. What is the semiconductor type expected ?

[*Ans.* .0358 eV and .0698 *e*V]

2.10. Elements of valency 6 have two surplus electrons over those needed to form covalent bonds. Find the ionisation energies of these electrons in Si. How does the impurity center behave if the ionisation energy is nearly half of the band gap energy ?

[*Ans.* .01286 *e*V]

2.11. A 12 mil thick Si Hall probe carrying a current of 3 mA develops a voltage of 1 mv at 2 K gauss. Find the Hall constant and doping level.

[*Ans.* $R = 1.016 \times 10^{+3}$ cm^3/coul
$p = 6.15 \times 10^{15}$/cm^3]

2.12. At a given temperature for what doping level and what type will a semiconductor show highest resistivity ?

[*Ans.* *p*-type doping, $p = n_i \sqrt{\mu_n/\mu_p}$.

2.13. The mobility for nearly intrinsic materials of Si, Ge and GaAs are proportional inversely to T^n where $n = 2.5$, 1.66 and 1 for electrons and 2.7, 2.33 and 2.1 for holes respectively. Find the temperatures at which the Hall constant is zero in these semiconductors and at what doping levels? [Assume $\mu = \mu_0 \left(\frac{300}{T}\right)^n$ where μ_0 for electrons and holes in cm^2/V sec 1500, 600 for Si, 3900 and 1900 for Ge and 8500 and 400 for GaAs respectively.]

[*Ans.* Si = 3°K, Ge = 102.56°K, GaAs = 18.64°K]

2.14. Find the absorption limit wave lengths for Si, Ge and GaAs. (Use room temperature band gaps as 1.12, 0.66 and 1.43 and $Eg = \hbar w$. Check that this gives Eg (eV) = (1.242/wavelength in microns)).

[*Ans.* Ge : 1.897 microns, Si : 1.109 microns, GaAs : 0.8692 microns]

2.15. Application of a d.c. bias to a semiconductor causes the electrons to oscillate in the E-K space. Oscillation in E-K space has corresponding oscillation in physical space. Would one expect oscillating charges to appear at the electrodes? Clarify with supporting numerical values. (Assume the conduction band energy spread a few or less electron volt and a mean free path. Then find for this energy change the order of magnitude of the field needed, which again would not take place because of collisions.

[*Ans.* No]

2.16. Argue in view of the answer to problem 5, Ch. 1 that the etch rate in the (111) plane in Si is expected to be lower than in (100) plane by a factor of at least 15 percent. What consequences is expected if through an window, in the (100) plane Si is etched?

[*Ans.* A V-groove bounded by (111) planes.]

2.17. Draw energy versus population density and the location of the Fermi energy in the following cases at T = 0°K.

(*i*) A monovalent metal, (*ii*) a divalent metal having good conductivity, (*iii*) a divalent metal having poor conductivity, (*iv*) a semiconductor, (*v*) a good insulator. Comment on their temperature coefficient of resistance.

2.18. For a trap level of density N_t deep within the energy gap show that there can be four processes—two for capture of electrons by unoccupied states and two for releasing of electrons by occupied states. Identify each process. Draw a diagram representing these four processes. Trap energy level is E_t.

2.19. Show that the rate of capture of electrons by unoccupied states from the conduction band in the above problem under non-equilibrium conditions is given by

$$C_n(E)\, N_t.(1-f_t)\, f_c(E).N_c(E)dE$$

where $f_t(E_t) = [1 + \exp(E_t - F_t)/kT]^{-1}$, F_t being the quasi-Fermi level for traps, and $f_c(E) = [1 + \exp(E - E_{Fn})/kT]^{-1}$ are the Fermi functions of traps and the conduction band respectively, while $N_c(E)$ is the conduction band seat numbers for unit energy range. Note that under non-equilibrium situations F_t and E_{Fn} and not the same Fermi levels.

2.20. Show that the occupied states of problem (2.18) above release their electrons to the conduction band at a rate

$$e_n(E) N_t f_t N_c(E)(1 - f_c)\, dE$$

where $e_n(E)$ is the corresponding rate constant.

Hence find the net rate of capture of electrons from the conduction band. Equate this net rate to zero to satisfy the principle of detail balancing under equilibrium conditions and since in equilibrium the same Fermi level is involved, find that the ratio

$(e_n/c_n) = [\,(f_c/1 - f_c)][(1 - f_t)/f_t]$ evaluated at equilibrium and is equal to $[\,\exp - (E - E_t)/kT\,]$.

2.21. Show that under non-equilibrium conditions the net rate of capture of electrons summed over all energies above the conduction band can be written as

$$R_n = [\,N_t(1 - f_t) - N_t f_t \exp - (E_{Fn} - E_t)/kT\,]\, nC_n$$

where

$nC_n = \displaystyle\int_E^\infty C_n(E) N_c(E) f_c(E) dE$ and check that

$$n_t = n \exp\,[-(E_{Fn} - E_t)/kT\,] = N_c \exp\left(-\frac{E_c - E_t}{kT}\right)$$

where n_t is the concentration of conduction band electrons if the Fermilevel coincide with trap energy.

2.22. Show that the net rate of electron capture obtained in problem (2.21) above can be written as

$$R_n = C_n N_t(1-f_t)n - C_n N_t f_t n_t$$

where $n_t = N_c \exp[-(E_c - E_t)/kT]$. Hence find a physical meaning of n_t by writing in the form $n_i \exp-(E_i - E_t)/kT$.

2.23. Treat the case of hole capture by the trap level along the lines indicated in problems (2.19) to (2.22) and obtain for holes

$$R_p = C_p N_t f_t p - C_p N_t(1-f_t)p_t \text{ where}$$

$$p_t = N_v \exp[-(E_t - E_v)/kT] = n_i \exp(E_i - E_t)/kT.$$

2.24. Find the value of $f(t)$ under steady state, equating $R_n = R_p$ of problems (2.22) and (2.23) above and show that the recombination rate

$$R = (pn - p_t n_t)/[\,T_n(p+p_t) + T_p(n+n_t)\,] \text{ where}$$

$$T_n = 1/C_n N_t \text{ and } T_p = 1/C_p N_t.$$

Equating $R_n = R_p$ find f_t substituting the value of which in R_n or R_p ; one has R.

2.25. Show that the product of p_t and n_t of problems (2.22) and (2.23) above is equal to the square of the intrinsic concentration. Note therefore $p_t n_t = p_0 n_0$ where p_0 and n_0 are equilibrium hole and electron concentrations in the respective bands.

2.26. How can the effective masses be determined experimentally ?

Ans. From Cyclotron resonance experimental results.

Chapter 3

HOMO–JUNCTIONS

In this chapter we shall discuss about junctions. Junctions can be classified as of two types: (1) the metal to semiconductor junctions, and (2) the semiconductor to semiconductor junction. The latter type can again be divided into two categories: (*i*) Junctions between P and N type semiconductors of the same material; (*ii*) Junctions of semiconductors having different band-gap. In the following we shall discuss the junctions between P and N type semiconductors of the same band-gap material.

3.1 P-N Junctions

These junctions of P and N types of the same material can also be of two types : (*i*) the abrupt junction, in which the impurity type changes abruptly at a plane within the bulk semiconductor (*ii*) the grown junction or graded junction in which the impurity type changes linearly through zero concentration from one type to the other. Here we shall discuss the first one and draw a few important conclusions regarding the properties of the step-junction, so called because the impurity type changes in a step as shown in Fig. (3.1.1)

Let us assume that a P-N junction has been formed of Ge material. If the doping of the P-side is 10^{16}/cc and the N-side 10^{15}/cc then in the bulk of the semiconductors, in the P-type there will have 10^{16} holes and 10^{10} electrons since in Ge $n_i = 10^{13}$/cc at room temperature while in the N-type bulk there should have 10^{15}/cc electrons and 10^{11}/cc holes. Thus we see at the junction region from P to N side there is a concentration gradient of holes from 10^{16}/cc to 10^{11}/cc and of electrons from 10^{10}/cc to 10^{15}/cc. Hence near the junction holes will diffuse from P-type to N-type sample and electrons from N-type to P-type. As the holes move from P-type, they leave bare immobile negative charges

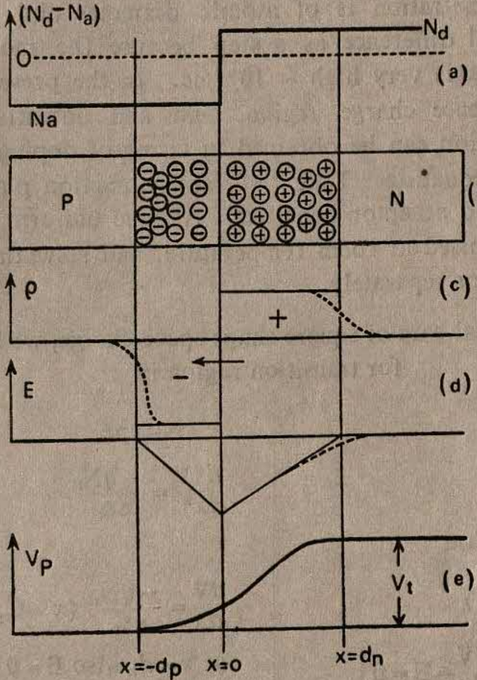


Fig. 3.1.1. (a) Qualitative picture of doping concentration ; (b) Space charges (ρ) ; (c) Shows space charge density ; (d) Electric field ideal (bold line) and actual cases (dotted) around the Junction ; (e) The potential variation

while electrons moving from N-type leave fixed positive charges behind. Thus a redistribution of mobile carriers takes place, and an electric field is established, [see Fig. (3.1.1d) directed from N- to P-type]. A dynamic equilibrium is reached in a very short time such that the field current of the same carrier type balances the diffusion current in the charge depleted region (the transition region or the space charge region). The field becomes that much strong as is needed to return the same number of diffusing holes or electrons at each moment that are crossing the junction. This readjustment of carrier distribution is confined within a very small distance around the junction. For a certain number of lines
]

of force, greater the doping density less will be the distance of penetration of them to obtain the required bare charges for their termination. The extreme case is metal to metal contact, where the depth of penetration is of atomic diameter and we get the contact potential difference as a step because the mobile carrier density in metals is very high ~ 10^{28}/cc. In the present case the width of the space charge region, field and potential variation around the junction can be obtained in terms of doping density by using Poisson's equation. Take $x=0$, at the junction plane, N_d and N_a the donor and acceptor densities. Assume uniform doping and all atoms are ionised at room temperature, and solve the equations for the two regions separately.

In the ideal case of square shape space-charge, we have for transition region in

P-type	N-type	
$\frac{d^2V}{dx^2}=\frac{+qN_a}{\varepsilon\varepsilon_0}$	$\frac{d^2V}{dx^2}=\frac{-qN_d}{\varepsilon\varepsilon_0}$	... (3.1.1)
gives on integration		
$\frac{dV}{dx}=\frac{qN_a}{\varepsilon\varepsilon_0}(x+C_1)$	$\frac{dV}{dx}=\frac{-qN_d}{\varepsilon\varepsilon_0}(x+C_2)$	... (3.1.2)
at $x=-d_p$ $\left(-\frac{dV}{dx}=E=0\right)$	at $x=d_n$ also $E=0$	
$\frac{dV}{dx}=\frac{qN_a}{\varepsilon\varepsilon_0}(x+d_p)$	$\frac{dV}{dx}=\frac{-qN_d}{\varepsilon\varepsilon_0}(x-d_n)$	... (3.1.3)

Further integration gives with C the potential at $x=0$

$$V=\frac{qN_a}{\varepsilon\varepsilon_0}\left(\frac{x^2}{2}+xd_p\right)+C \text{ and } V=\frac{-qN_d}{\varepsilon\varepsilon_0}\left(\frac{x^2}{2}-xd_n\right)+C.$$

The voltage at $x=-d_p$ is $V_p=C-\frac{qN_a}{2\varepsilon\varepsilon_0}d_p^2$

while at $x=d_n$ is $V_n=C+\frac{qN_d}{2\varepsilon\varepsilon_0}d_n^2$.

Hence the transition region voltage

$$V_t=V_n-V_p=\frac{q}{2\varepsilon\varepsilon_0}(N_a d_p^2+N_d d_n^2) \qquad \cdots \text{(3.1.3a)}$$

We also have for overall electrical neutrality, total exposed charge on P-side equal to that on N-side for each unit area

$$qN_d d_n = qN_a d_p \text{ that is } \frac{d_n}{d_p} = \frac{N_a}{N_d} \qquad \cdots \quad (3.1.4)$$

From (3.1.3a) and (3.1.4) we have

$$d_n = \left[\frac{2\varepsilon\varepsilon_0}{q} V_t \frac{N_a}{N_d(N_a+N_d)}\right]^{\frac{1}{2}} \text{ and}$$

$$d_p = \left[\frac{2\varepsilon\varepsilon_0}{q} V_t \frac{N_d}{N_a(N_a+N_d)}\right]^{\frac{1}{2}} \qquad \cdots \quad (3.1.5)$$

Hence the total depletion region width is

$$d = d_n + d_p = \left[\frac{2\varepsilon\varepsilon_0}{q} . V_t \left(\frac{1}{N_a} + \frac{1}{N_d}\right)\right]^{\frac{1}{2}}$$

$$= \left[\frac{2\varepsilon\varepsilon_0}{q} \frac{V_t}{N}\right]^{\frac{1}{2}} \qquad \cdots \quad (3.1.6)$$

where N, the effective doping density is expressed as

$$\frac{1}{N} = \frac{1}{N_a} + \frac{1}{N_d} \qquad \cdots \quad (3.1.7)$$

The depletion region will act as a capacitor. A small forced change in V_t would cause a rearrangement of some mobile carriers on both sides, the net result being equal total fixed exposed charge Q on both sides of the junction. Thus junction capacitance

$$C_j = \frac{dQ}{dV_t} = \frac{d}{dV_t}(AqN_d d_n)$$

$$= A\sqrt{\frac{\varepsilon\varepsilon_0}{2} \frac{qN}{V_t}} = \frac{A\varepsilon\varepsilon_0}{d} \qquad \cdots \quad (3.1.8)$$

To calculate the built in potential $V_b = V_t$ (under equilibrium conditions) one can start from the current expressions.

$$J_n = q\,\mu_n n E + D_n\, q \frac{dn}{dx} \qquad \cdots \quad (3.1.9)$$

$$J_p = q\,\mu_p p E - D_p\, q \frac{dp}{dx} \qquad \cdots \quad (3.1.10)$$

which equated to zero gives for the field

$$\mathrm{E} = -\frac{\mathrm{KT}}{q}\cdot\frac{1}{\bar{n}}\cdot\frac{d\bar{n}}{dx} = \frac{k\mathrm{T}}{q}\cdot\frac{1}{\bar{p}}\cdot\frac{d\bar{p}}{dx} \qquad \cdots \quad (3.1.11)$$

bar and knot standing for thermal equilibrium and depletion boundary values respectively. Using (3.1.11) one finds

$$\mathrm{V}_b = \int_{x=-d_p}^{x=d_n} -\mathrm{E}dx = \frac{k\mathrm{T}}{q}\int_{n_{p_0}}^{n_{n0}} \frac{1}{\bar{n}}\,\frac{d\bar{n}}{dx} = -\frac{k\mathrm{T}}{q}\Big[\log p\Big]_{p_{p_0}}^{\bar{p}_{n0}} = \frac{k\mathrm{T}}{q}\log\frac{\mathrm{N}_d\mathrm{N}_a}{n_i^{2}}$$

$$\cdots \quad (3.1.12)$$

For example, $\mathrm{N}_a = 10^{16}$/cc and $\mathrm{N}_d = 10^{15}$/cc give $\mathrm{V}_b = 240$ mV for Ge and 620 mV for Si at room temperature (300°K).

Note that for a step junction the field is negative in the transition region and maximum at the junction. The depletion region width varies directly as the square root of the voltage V_t and inversely as the square root of the effective doping density. If $\mathrm{N}_a \gg \mathrm{N}_d$, then $\mathrm{N} = \mathrm{N}_d$ and $d \propto \sqrt{\mathrm{V}_t/\mathrm{N}_d}$ that is the depletion region width is determined by the doping concentration of lightly doped material and extends mostly in the lightly doped material (check $d_p \ll d_n$).

For an estimation of the width of the depletion region, use $q = 1{\cdot}6 \times 10^{-19}$ coul, $\varepsilon_0 = 8{\cdot}85 \times 10^{-14}$ F/cm. $\varepsilon = 16$ for Ge and 12 for Si. For Ge this gives $d = 0{\cdot}64$ microns with $\mathrm{V}_b = 240$ mV. Hence when $\mathrm{V}_t = 24$ volts, the depletion region width will be 6·4 microns. Prebias field is $6{\cdot}5 \times 10^{3}$ V/cm at the junction boundary in the former case while in the latter $6{\cdot}5 \times 10^{4}$ V/cm.

It can be shown for a graded (grown) junction with doping density varying linearly like bx where b is a positive constant, d is given by

$$d = \left[\frac{12\,\varepsilon\varepsilon_0}{qb}\right]^{1/3} \text{and } \mathrm{C}_j = \mathrm{A}\left[\frac{qb\,\varepsilon^2\varepsilon_0^2}{12\mathrm{V}_t}\right] = \frac{\mathrm{A}\varepsilon\varepsilon_0}{d} \qquad \cdots \quad (3.1.13)$$

In this case the starting equations on either side of $x = 0$ are,

$$\frac{d^2\mathrm{V}}{dx^2} = \frac{-qbx}{\varepsilon\varepsilon_0} \text{ for } x \leqslant 0 \qquad \cdots \quad (3.1.14)$$

(x is negative in this eqn.)

and $$\frac{d^2\mathrm{V}}{dx^2} = \frac{-qbx}{\varepsilon\varepsilon_0} \text{ for } x \geqslant 0$$

for P- and N-type respectively while overall charge neutrality requires $d_n^2 = d_p^2$ as obtained from

$$q\int_0^{d_n} N_d\,dx = q\int_0^{d_p} N_a\,dx$$

with $N_d = N_a = bx$.

Check that the maximum field occurs at $x=0$ and the relationship this E_{max} bears to the voltage across the transition region V_t is

$$V_t = \frac{4}{3}\left(\frac{2}{b}\,\frac{\varepsilon\varepsilon_0}{q}\right)^{1/2} E_m^{3/2} \qquad \cdots \quad (3.1.15)$$

3.2 P-N Junction from Energy band approach

In this article we consider the p-n junction of uniformly doped P & N materials brought into contact. The assumptions are, that the crystal structure is continuous and ideally clean at the junction (single crystal).

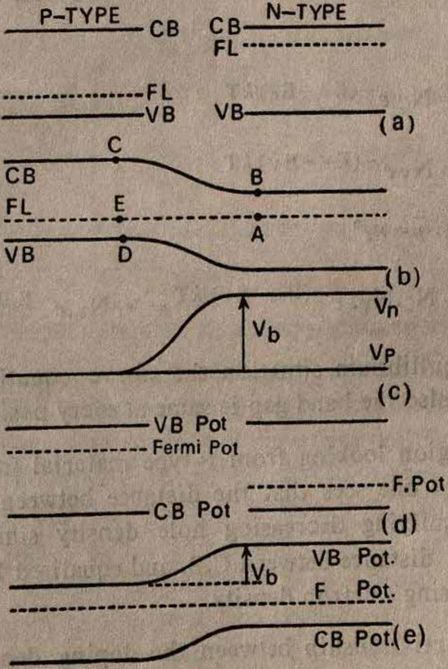


Fig. 3.2.1. Shows the P-N Junction from the energy and potential approach respectively

Before contact F. L. in P-type material remains near the valence band while F. L. in N-type material is near the conduction band as shown in Figure 3.2.1. Remembering that the energy band diagram depicts electron energy, the corresponding potential picture would be, energy divided by negative charge, hence a reversal as shown in Fig. (3.2.1d) and (3.2.1e). Before contact the holes in P-type are at higher potential so at higher energy while electrons in N-type are at lower potential and hence at higher energy, therefore on contact holes will roll down the potential gradient at the contact until the Fermi-potential is equalised. Thus a dynamic equilibrium is reached by a relative upward shift of V. B. potential & C. B. potential of the N-type material with respect to those of the P-type material. This results in a potential rise from the P-type to N-type material in the transition region as shown in Fig. (3.2.1c).

As for the concentration of carriers in the transition region under equilibrium conditions, the electron and hole densities will be given by

$$n = \mathrm{N_C}\, e^{-(\mathrm{E_C} - \mathrm{E_F})/k\mathrm{T}} \quad \cdots \quad (3.2.1)$$

$$\bar{p} = \mathrm{N_V}\, e^{-(\mathrm{E_F} - \mathrm{E_V})/k\mathrm{T}} \quad \cdots \quad (3.2.2)$$

$$\bar{n}\,\bar{p} = n_i^2$$

$$= \mathrm{N_C}.\mathrm{N_V}.\, e^{-(\mathrm{E_C} - \mathrm{E_V})/k\mathrm{T}} = \mathrm{N_C}\mathrm{N_V}.e^{-\mathrm{E}_g/k\mathrm{T}} \quad \cdots \quad (3.2.3)$$

In this equilibrium situation the above equations are everywhere satisfied, also the band gap is same at every position.

In this region looking from P-type material (see Fig. 3.2.1b) towards the right one sees that the distance between the FL & VB is increasing signifying decreasing hole density similarly from N-type to the left, distance between C.B. and equalised F.L. increasing, signifying decreasing electron density.

To find a relationship between the doping densities, built in potential V_b and band gap E_g, one can sum up the electron energies in Fig. (3.2.1b).

A to B $= -k\mathrm{T}\log\frac{n_n}{\mathrm{N}_c} = (\mathrm{E}_c - \mathrm{E}_f)$ in N-type a rise of energy, where n_n is the electron density in N-type bulk

B to C $= q\mathrm{V}_b = -q\,(-\mathrm{V}_b)$ a rise of energy

C to D $= -q\mathrm{V}_g = -\mathrm{E}_g$ a fall of energy

D to E $= -k\mathrm{T}\,\frac{p_p}{\mathrm{N}_v} = (\mathrm{E}_f - \mathrm{E}_v)$ in P-type a rise of energy where p_p is the hole density in the p-type bulk, giving us

$$q\mathrm{V}_b - q\mathrm{V}_g - k\mathrm{T}\,\log\frac{p_p}{\mathrm{N}_c}\frac{n_n}{\mathrm{N}_v} = 0 \qquad \cdots \quad (3.2.4)$$

$$\text{or}\quad \mathrm{V}_b = \mathrm{V}_g + \frac{k\mathrm{T}}{q}\,\log\frac{p_p\;n_n}{\mathrm{N}_c\,\mathrm{N}_v} \qquad \cdots \quad (3.2.5)$$

Note that $\mathrm{V}_b < \mathrm{V}_g$ since $p_n \leqslant \mathrm{N}_v$ & $n_n \leqslant \mathrm{N}_c$ also assuming dopant atoms are all ionised we have

$$\mathrm{V}_b = \mathrm{V}_g + \frac{k\mathrm{T}}{q}\,\log\frac{\mathrm{N}_a\,\mathrm{N}_d}{\mathrm{N}_c\,\mathrm{N}_v} \qquad \cdots \quad (3.2.6)$$

substituting (3.2.3) in (3.2.5) with $\mathrm{E}_g = q\mathrm{V}_g$ we have

$$\mathrm{V}_b = \mathrm{V}_g + \frac{k\mathrm{T}}{q}\,\log\frac{p_p\,n_i^2}{p_{no}\mathrm{N}_c\mathrm{N}_v} = \mathrm{V}_g + \frac{k\mathrm{T}}{q}\,(\log\frac{p_p}{p_{no}} + \log e^{-q\mathrm{V}_g/k\mathrm{T}})$$

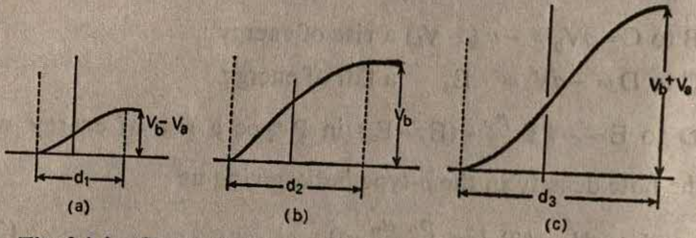
$$= \frac{k\mathrm{T}}{q}\,\log\frac{p_p}{p_{no}} = \frac{k\mathrm{T}}{q}\,\log\frac{n_n}{n_{po}} \qquad \cdots \quad (3.2.7)$$

which is the same as eqn. (3.1.12) since $p_p = p_{po}$ and $n_n = n_{no}$ at the depletion region boundaries.

3.3 Minority Injection and boundary concentrations

The electrical resistance of doped medium being very low we assume, when bias is applied to a p-n junction diode, the rest of the bulk, than the junction, does not have any voltage drop. Therefore the applied voltage drops within the junction region only. Under reverse bias of applied voltage V_a, the N-material is positive with respect to the P-region, hence the previously existing field increases causing the voltage drop across the junction to increase to $\mathrm{V}_t = \mathrm{V}_b + \mathrm{V}_a$, with the consequent widening of the transition region width to that given by eqn. (3.1.6). For forward bias the converse is the case and shown in (Fig. 3.3.1). Note that $d_3 > d_2 > d_1$.

Ignoring the boundary shift, say under forward biased condition, p_{no} and n_{po} are the minority carrier concentrations at



(a)
(b)
(c)

Fig. 3.3.1. Shows the junction region potential for the three cases (a) Forward, (b) Pre-bias, (c) Reverse bias.

N and P region boundaries respectively. Then we have from equation (3.1.12)

$$(V_b - V_a) = (kT/q) \log \frac{p_p}{p_{no}} = \frac{kT}{q} \log \frac{n_n}{n_{po}}$$

where $p_p = p_{po}$ and $n_n = n_{no}$ are assumed, which with eqn. (3.2.7), can be written as

$$p_{no} = p_p\, e^{-q(V_b - V_a)/kT} = \bar{p}_{no}\, e^{qV_a/kT} \qquad \cdots \quad (3.3.1)$$

The previous boundary concentrations were $\bar{p}_{no}$ and $\bar{n}_{po}$ respectively. Hence increase in boundary concentrations are

$$\left.\begin{aligned} \hat{p}_o &= \bar{p}_{no}\,[\exp(qVa/kT) - 1] \text{ at N-material boundary} \\ \hat{n}_o &= \bar{n}_{po}\,[\exp(qVa/kT) - 1] \text{ at P-material boundary} \end{aligned}\right\} \quad \cdots \quad (3.3.2)$$

Hence we speak of injection of minority carriers in the samples.

Note that for a reverse bias as small 100 mV the injected concentration $\hat{p}_o$ tends to $-\bar{p}_{n}$, for since at room temperature $q/kT \simeq 40$ per volt. Hence even for small reverse bias the boundary concentration tends to zero.

If the applied voltage is a d.c. plus a superposed a.c. like $Va = Ve + \overset{\leftharpoondown}{v}.e^{j\omega t}$, then the excess concentration at the two boundaries will have d.c. and a.c. parts. Using symbols $\hat{p}_o$ for the d.c. part and $\overset{\leftharpoondown}{p}_o$ for the amplitude of the a.c. part, for holes from equation (3.3.2) we have

$$\hat{p}_o + \overset{\leftharpoondown}{p}_o + \bar{p}_{no} = \bar{p}_{no} \exp(qVe/kT).\ \exp(q\overset{\leftharpoondown}{v}/kT) \qquad \cdots \quad (3.3.3)$$

From (3.3.3) for $\overleftharpoon{v}=0$, we have

$$\hat{p}_o+\bar{p}_{no}=\bar{p}_{no}\cdot\exp\left(qVe/kT\right) \quad \cdots \quad (3.3.4)$$

From (3.3.3) and (3.3.4)

$$\hat{p}_o+\overleftharpoon{p}_o+\bar{p}_{no}=\left(\hat{p}_o+\bar{p}_{no}\right).\exp\left(q\overleftharpoon{v}/kT\right) \quad \cdots \quad (3.3.5)$$

For $\overleftharpoon{v}\leqslant\dfrac{kT}{q}$ (small signal) the exponential can be expanded in a series and neglecting higher order terms we have

$$\hat{p}_o+\overleftharpoon{p}_o+\bar{p}_{no}=\left(\hat{p}_o+\bar{p}_{no}\right)^{1}\left(+\frac{q\overleftharpoon{v}}{kT}\right) \quad \cdots \quad (3.3.6)$$

hence $\overleftharpoon{p}_o=\left(\bar{p}_{no}+\hat{p}_o\right)\dfrac{q\overleftharpoon{v}}{kT}$ $\quad \cdots \quad (3.3.7)$

Injected d.c. minority carrier level $\hat{p}_o$ is normally very high compared to the thermal equillibrium level $\bar{p}_{no}$ in forward bias. Therefore, we can write

$$\overleftharpoon{p}_o=\frac{q}{kT}\cdot\overleftharpoon{v}\cdot\hat{p}_o \quad \cdots \quad (3.3.8)$$

Similarly, at the boundary in the P-type the excess minority concentration

$$\overleftharpoon{n}_o=\frac{q}{kT}\cdot\overleftharpoon{v}.\hat{n}_o \quad \cdots \quad (3.3.9)$$

and

$$\left.\begin{aligned}\hat{p}_o&=\bar{p}_{no}\left[\exp\left(qVe/kT\right)-1\right]\text{ at N-boundary}\\ \hat{n}_o&=\bar{n}_{po}\left[\exp\left(qVe/kT\right)-1\right]\text{ at P-boundary}\end{aligned}\right\} \quad \cdots \quad (3.3.10)$$

Equation (3.3.10) describes the d.c. while (3.3.8) and (3.3.9) describe the a.c. boundary concentrations of minority carriers related to the forward biased junction. For a reversed biased junction the boundary concentrations are zero. In a forward biased junction the variation of the depletion region width may be neglected but in reverse bias this may be appreciable.

The above equation indicates that it is possible to have a single type of carrier diode, for example, assuming all donors to be ionised the minority injection levels on either side of the boundary can be written from (3.3.10) as

$$\left.\begin{aligned}\hat{p}_o &= \frac{n_i^2}{N_d}\,[\exp(qVe/kT)-1] \text{ on N-boundary and}\\ \hat{n}_o &= \frac{n_i^2}{N_a}\,[\exp(qVe/kT)-1] \text{ on P-boundary}\end{aligned}\right\} \quad \cdots \quad (3.3.11)$$

So, if the P-material is much more heavily doped than the N-material *i.e.*, $N_a \gg N_d$ then the excess minority level $\hat{n}_o$ on P-boundary is very much smaller than $\hat{p}_o$ on N-boundary. Thus the contribution to the total current by injection in P-type is much smaller than due to the hole injection in N-material. Thus we have a hole carrier diode. Alternatively by making $N_d \gg N_a$ we can have an electron carrier diode.

3.4 Ohmic Contact

An ohmic contact for a particular type of carrier is one in which this carrier can easily flow out while the other type of carrier can flow in through the same contact such that the current voltage relationship of the contact region follows the Ohm's law. That is the contact resistance is independent of the direction of current flow. This can be obtained if there is a suitable potential distribution between the semiconductor and the metal contact.

The boundary concentration for such a contact cannot build up hence it is same as the thermal equilibrium concentration. That is, the excess part of minority carrier concentration at the ohmic contact used to drain this excess will be taken as zero. Such a contact can be obtained if the metal being used for forming the contact by alloying, inter diffusion, coating or plating is itself a dopant impurity of the same type as the semiconductor. Care, of course, has to be taken adequately to have clear surfaces for making ohmic contact.

3.5 Carrier decay processes

In this article we investigate, whether it is possible to build-up, maintain and control majority carrier concentration. Consider that an N-doped ($N_d = 10^{16}$/cc) germanium crystal with alloyed ohmic contacts is biased as shown in Fig. (3.5.1). Let within the body of the semiconductor at $x=0$ local momentary majority concentration disturbance of $n_o=10^{14}$/cc (only 1% of the existing electron concentration, all donors assumed ionised) has been introduced.

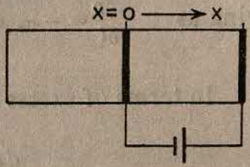


Fig. 3.5.1. Showing the injection point by application of bias

These additionally introduced electrons form a momentary space charge and we make a rough estimate of the space charge field involved by integration of Poisson's eqn.

$$\frac{d^2V}{dx^2}=\frac{10^{14}q}{\varepsilon\varepsilon_o} \qquad \cdots \quad (3.5.1)$$

On integration with C the const. of integration

$$\frac{dV}{dx}=\frac{10^{14}qx}{\varepsilon\varepsilon_o}+C \qquad \cdots \quad (3.5.2)$$

assuming at a distance of 1 mm. on the left of $x=0$ the field is zero, we evaluate the constant C and from (3.5.2) write for the field

$$E=-\frac{10^{14}q}{\varepsilon\varepsilon_0}(x+0{\cdot}1) \qquad \cdots \quad (3.5.3)$$

from which the field at $x=0$ is $E=\frac{-10^{13}q}{\varepsilon\varepsilon_0}\simeq-10^6$V/cm

which is very much larger than the breakdown field in semiconductors ($\sim 10^5$V/cm). Though a very inaccurate estimation of the field, it is sufficient to indicate the impossibility of maintaining space charge non-neutrality and that the field would sweep the excess majority carriers to the collecting terminal on the right.

Now, let us examine the time involved in clearing or relaxing this space charge.

For this we have the continuity eqn.

$$\frac{\delta\rho}{\delta t}=-\text{div. J} \qquad \cdots \quad *(3.5.4)$$

---

* Equation (3.5.4) can be established easily as follows : Consider a small volume A. δx of the material. If the current decreases by δi while passing through the length δx then the charge loss $-\delta i.\ \delta t$ in a small time δt is equal to that in building up the charge of $\delta\rho$. A. δx in volume Aδx. Equating we have $\frac{\delta\rho}{\delta t}=-\frac{1}{A}\cdot\frac{\delta i}{\delta x}=\frac{-\delta J}{\delta x}$ in the limit we have (3.5.4).

and current density $J = \sigma E$ $\cdots$ (3.5.5)

giving us $$\frac{\delta \rho}{\delta t} = -\sigma .\ \text{div. } E = -\frac{\sigma \rho}{\varepsilon \varepsilon_o} \qquad \cdots \quad (3.5.6)$$

In terms of excess carrier density $\hat{n}$ write (3.5.6) as

$$\frac{d\hat{n}}{dt} = -\sigma \frac{\hat{n}}{\varepsilon \varepsilon_o} \qquad \cdots \quad (3.5.7)$$

The solution of which is

$$\hat{n} = \hat{n}_o \exp\left(-t / \frac{\varepsilon \varepsilon_o}{\sigma}\right) \qquad \cdots \quad (3.5.8)$$

The excess majority carrier concentration decays exponentially with a time constant $T_{relax} = \frac{\varepsilon \varepsilon_o}{\sigma} \sim 10^{-13}$ secs. Note that

$$\sigma = q(\mu_n n + \mu_p p) = \mu_n q(\overline{n} + \hat{n}) = 3 \cdot 6 \times 10^3 \times 1 \cdot 6 \times 10^{19} \times 1 \cdot 01 \times 10^{16}$$

$$= 56 \cdot 34 \ \frac{\text{mho}}{\text{cm}}$$

is determined by not the excess carriers alone but mostly by the previously existing concentration of 10^{16}/cc. Therefore, the majority carrier charge relaxation takes a very short time because it is not the excess concentration alone that is responsible for their removal but in the very short duration of current flow, all the majority carriers take part. The conductivity, or that, the majority concentration is high making T_{relax} short.

The above consideration shows that use of majority carrier control in a device is difficult unless the device dimension or the path resistance is controlled.

The situation, however, is vastly different if instead of majority carriers the same number of minority carriers are introduced at the same place in our sample. Firstly, the order of magnitude of the transient field is the same but the direction is opposite. Secondly, the concentration of holes that take part in this decay process is $\left(10^{14} + \frac{n_i^2}{10^{16}}\right) \simeq 10^{14}$/cc. This decay of the excess minority carriers is much slower therefore, than the build up of majority carriers. Hence this decay does not take place immediately at all, instead the majority carriers are drawn in by this transient space charge field until the space charge field is quenched.

However, this condition does not persist indefinitely either. In an unbiased semiconductor the holes and electrons remain in the same medium and they maintain their concentrations. The state is a dynamic equilibrium where constantly there is generation and recombination of carriers but maintaining always $\bar{n}\,\bar{p}=n_i^2$. If the generation rate is g and the recombination constant is r then we may write the time rate change for small deviations from equilibrium, as

$$\frac{dn}{dt}=\frac{dp}{dt}=g-r.n.p \qquad \cdots \quad (3.5.9)$$

Under equilibrium condition time rate change of concentration is zero hence

$g=r\,\bar{n}\,\bar{p}=r.\,n_i^2$ when (3.5.9) can be written as

$$\frac{dp}{dt}=r.\,(n_i^2-np) \qquad \cdots \quad (3.5.10)$$

If minority concentration changes by $\hat{p}$ for charge neutrality, majority concentration also changes by $\hat{n}=\hat{p}$ hence from (3.5.10), we have since $p=\bar{p}+\hat{p}$

$$\frac{dp}{dt}=r[n_i^2-(\bar{n}+\hat{n})\ (\bar{p}+\hat{p})]\simeq -r\,p\,(\bar{n}+\hat{n})$$

$$=-\frac{p}{T_p} \qquad \cdots \quad (3.5.11)$$

equation (3.5.11) predicts that the excess minority carriers (here holes) decay exponentially with a time constant T_p where $T_p=\dfrac{1}{r(\bar{n}+\hat{n})}$, the life time of holes and is determined by the recombination process and the majority carrier density. Note even the neutral deviation $\hat{p}=\hat{n}$ from thermal equilibrium does not last for ever, but decays remaining neutral till the decay is complete, with the time constant T_p exponentially.

Similarly, if the sample was P-type and minority carriers, electrons were introduced then they would decay exponentially with a life time of electrons. The minority life times are largely dependent on crystal perfection. More the imperfection is less is the minority life time. These imperfections act as recombination centers hence increase r thus decreasing the life time. Very good single crystals show minority life times as high as 10^{-3} secs, while controlled imperfection can create minority life time as low as 5×10^{-8} secs. In good single crystals of semiconductors the minority life times (holes in N-type and electrons in P-type) are much larger than the majority relaxation times and that in both cases it is the majority carriers that control the decay process (higher the majority carrier density lower is the minority life time). If the concentration disturbance is a majority carrier alone, it dies with majority relaxation T_{relax} while if it is a minority carrier disturbance then firstly majority carrier builds up to neutralise space charge and then both the majority and minority carriers decay simultaneously with the minority life time (T_p or T_n) as the case may be.

If there is no flow of minority carriers, then the loss of minority carriers can occur due to recombination alone while if there is recombination as well as current flow, then the minority carrier continuity equation can be written with the help of (3.5.4) and (3.5.11) for holes in N-type sample as,

$$\frac{\delta \hat{p}}{\delta t} = -\frac{1}{q}\frac{\delta J_p}{\delta x} - \frac{\hat{p}}{T_p} \qquad \cdots (3.5.12)$$

for electrons in P-type sample

$$\frac{\delta \hat{n}}{\delta t} = +\frac{1}{q}\frac{\delta J_n}{\delta x} - \frac{\hat{n}}{T_p} \qquad \cdots (3.5.13)$$

At this point the student is referred to read the celebrated Shockley Haynes experiment from (ref. 1 of ch. 12).

3.6 Determination of doping profile

It is known that if one side of a p-n junction is very heavily doped compared to the other side then the depletion region extends in the lightly doped region. The extension of the depletion in the heavily doped region can be easily ignored. For any arbitrary doping profile in the lightly doped region, as shown in Fig. (3.6.1), a small change in reverse bias voltage dV causes a change of depletion region width $d\delta$ and the exposed charge is

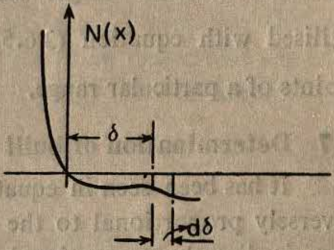


Fig. 3.6.1. Shows the impurity distribution for calculation of depletion width

$$Aq\,N(x)\,d\delta = C_j .\, dV \quad \cdots \quad (3.6.1)$$

$$\text{whereas } C_j = \frac{A\varepsilon\varepsilon_o}{\delta} \quad \cdots \quad (3.6.2)$$

Combining equations (3.6.1) and (3.6.2)

$$dV = \frac{A}{C_j}\,q\,N(x)\,d\left(\frac{A\varepsilon\varepsilon_o}{C_j}\right) = \frac{A^2 q\varepsilon\varepsilon_o N(x)}{2}\,d\left(\frac{1}{C_j^2}\right)$$

$$\text{Hence } N(x) = \frac{2}{A^2 q\varepsilon\varepsilon_o}\,\frac{dV}{d(1/C_j^2)} \quad \cdots \quad (3.6.3)$$

Hence plotting the reciprocal of the square of the junction capacitance versus reverse voltage curve can give us the doping profile from the above expression if the junction area A is known. It may be difficult to find the reciprocal slope $dV/d\left(\frac{1}{C_j^2}\right)$ for ample number of points of the curve. But since C_j would be a power law with voltage of the form $C_j = \frac{K}{(V_b+V_a)^n}$

$$\frac{1}{C_j^2} = \frac{(V_b+V_a)^{2n}}{K^2} \quad \cdots \quad (3.6.4)$$

$$\text{or} \quad \frac{d}{dV_a}\left(\frac{1}{C_j^2}\right) = \frac{2n(V_b+V_a)^{2n-1}}{K^2} = \frac{2n}{C_j^2(V_b+V_a)} \quad \cdots \quad (3.6.5)$$

For arbitrary doping profile, n will not, in general be constant for the whole range of voltage but this range can be divided into two or three small ranges in the $\log C_j$ vrs. $\log (V_b + V_a)$ curve since $\log C_j = \log K - n \log (V_b + V_a)$, to obtain different values of n for each of these ranges and each of these values of n can be utilised with equation (3.6.5) above to compute $\frac{d}{dV_a}(1/C_j^2)$ for points of a particular range.

3.7 Determination of built in voltage

It has been seen in equation (3.1.8) that junction capacitance is inversely proportional to the depletion region wtdth, which in its turn is directly proportional to a fractional power of the voltage across the transition region. This information can be utilised to find out experimentally the built in voltage of a junction.

The principle is to measure the junction reverse biased capacitance versus reverse bias voltage which is expected to be a monotonic curve as shown in Fig. (3.7.1). The total capacitance is composed of the header capacitance plus the junction capacitance. From the C vrs. V plot thus obtained, the C_j part can be picked up and replotted vrs. voltage. From the above expressions one has, if V_b is the built in voltage and V_a the applied reverse bias voltage,

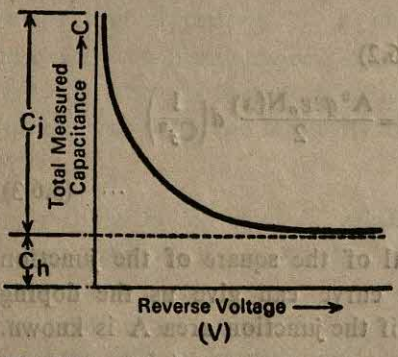


Fig. 3.7.1. Measured capacitance vers. reverse bias voltage plot

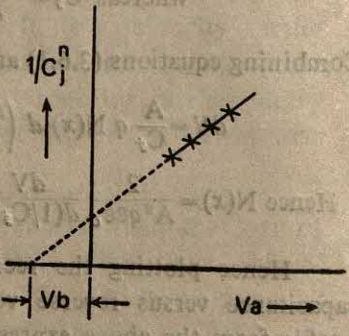


Fig. 3.7.2. Extrapolation of the linear portion of the curve giving built in voltage

$$C_j \propto \frac{1}{(V_b + V_a)^{1/2}} \text{ for step junction} \quad \cdots \quad (3.7.1)$$

$$\text{and } C_j \propto \frac{1}{(V_b + V_a)^{1/3}} \text{ for grown junction} \quad \cdots \quad (3.7.2)$$

So one can easily plot, $\frac{1}{C_j^2}$ vrs. V_a or $\frac{1}{C_j^3}$ vrs. V_a as the case may be a step junction or a grown one, to obtain points on the curve as shown in Fig. (3.7.2). These points can be joined and the curves can be extrapolated to intersect the voltage axis to obtain V_b as shown in Fig. (3.7.2)

$$\text{Since } (V_b + V_a) = \frac{\text{Const.}}{C_j^n} \qquad \cdots \quad (3.7.3)$$

i.e., if this law followed till zero transition region voltage, then the capacitance would have been infinite here, that is $1/C_j^n = 0$ when $V_b = -V_a$.

References

1. *Solid State Electronics*: S. Wang, McGraw Hill Book Company (Int. Students Edition, 1966)

2. *Fundamentals of Semiconductor Devices*: J. Lindmayer & C. Y. Wrigley, Van Nostrand Co. Inc. (1965)

Problems

3.1. A step junction diode is heavily doped with dopant concentration N_d on the n-side while p-side is intrinsic. Find the built in voltage of the junction.

Ans. $V_b = \frac{kT}{q} \log \frac{N_d}{n_i}$.

3.2. What is the built in voltage if the junction is a linearly doped one in problem 1.

Ans. Same as above.

3.3 For a p-n step junction diode of Si the n-side doping level is 10^{16} and the p-side doping is 10^{15}. Find the built in voltage, depletion region width and the capacitance of the junction. Assume a dielectric constant of 12 and a diameter of 100 microns.

Ans. $V_b = 631$ mV ; $d = {\cdot}95\ \mu m$; $C_j = 0{\cdot}9\ pF$.

3.4 Find the depletion region width and junction capacitance of a graded doped Si p-n junction diode with $N_d = 10^{20}x$, and

$N_a = 10^{19}x$ where x is in microns. Use $\varepsilon_r = 12$ and junction radius 25 microns.

Ans. 0·995 micron and 21 *p*F.

3.5. Find the depletion region width if the diode is forward biased with 0.5 volts applied voltage drop across the junction in the diode of problem (3.3). Find the forward biased junction capacitance and the minority carrier levels at the boundaries of the junction at room temperature.

Ans. $d = \cdot 437\ \mu m$; $C_j = 1{\cdot}9\ pF$; N-boundary $= 10{\cdot}9 \times 10^{12}$/cc ; $n = 10{\cdot}9 \times 10^{13}$/c.c.

3.6. If the junction of problem (3.3) is reverse biased by 5 volts, what is the depletion region width, the reverse biased capacitance and the minority carrier levels at the junction boundaries ?

Ans. $d = 2{\cdot}867\ \mu m$; $C_j = 0{\cdot}3\ pF$.

3.7. Calculate the shift of junction boundaries on the p- and n-sides of the problem (3.5) under forward bias. Find the excess carriers at these boundaries with the same bias.

Ans. Shift in N-side 0·047 μm
Shift in P-side 0·470 μm

3.8. Assuming our atmosphere consists of one species of gaseous particles and has a constant temperature T, show from gas laws and the definition of pressure that the gaseous particle concentration at a place is $n = n_0 \exp(-mgh/kT)$ where n_0 is the concentration and h the height from the sea level.

3.9. Find the temperature coefficient of the built in voltage of a Si diode and compare it with that of a Ge diode.

Hint : Use equation (3.2.5), find (dV_g/dt) from data of problem (2.7), Chapter II.

Ans. 2*m*V/°C nearly same for both cases.

3.10. Under the condition that the excess carrier concentrations n_1 and p_1 are small compared to their equilibrium values n_0 and p_0 show from the results of problems (2.24) and (2.25) of Ch. II, that single trap level controlled life time is given by

$$T = [T_n(p_0 + p_t) + T_p(n_0 + n_t)]/(n_0 + p_0)$$

Hint: Use $\frac{dp}{dt} = -R$ and proceed as given for equation (3.5.11). The student should note that the model used to arrive at the above result is known as the Schockley Read Model. The above formula is amply used for device performance calculation.

Chapter 4

P–N JUNCTION DIODE

We are thinking of injecting minority carriers into a semiconductor and their control while in transit through the sample like vacuum tubes, therefore, it is reasonable to suppose that lesser the transit time higher is the frequency at which the device would be usable. Therefore, we need to examine how the transit time can be minimised. The simple method of lowering the length of the medium being traversed by carriers is always used. On top of this for high frequency devices, it is customary to build a field in the medium which, in turn, increases the drift velocity of the injected minority carriers.

4.1 Built in field

This sort of building of a field in the semiconductor body is possible by doping the material with varying dopant density along the sample length. The standard techniques for achieving this are high temperature diffusion of dopant material in the sample or doping while growth occurs as in the epitaxial method.

To formulate the general relationship between space dependent doping density and potential variation, we start from the field expression at equilibrium

$$E = -\frac{kT}{q}\cdot\frac{1}{\bar{n}}\cdot\frac{d\bar{n}}{dx} = \frac{kT}{q}\,\frac{1}{\bar{p}}\cdot\frac{d\bar{p}}{dx} \qquad \cdots \quad (4.1.1)$$

while $V = -\int E\,dx = \frac{kT}{q}\log\frac{\bar{n}(x)}{\text{Const}} = -\frac{kT}{q}\log\frac{\bar{p}(x)}{\text{Const}} \qquad \cdots \quad (4.1.2)$

Choose the intrinsic concentration as the reference for voltage, which has not got to exist in the sample because the potential difference is what we are concerned with and not its absolute value ever, and write

$$\bar{n} = n_i . e^{qV/kT} \text{ and } \bar{p} = n_i . e^{-qV/kT} \qquad \cdots \quad (4.1.3)$$

Poisson's equation can be framed as

$$\frac{d^2V}{dx^2}=\frac{-q}{\varepsilon\varepsilon_0}[N_d+\bar{p}-\bar{n}] \qquad \cdots \quad (4.1.4)$$

which with equation (4.1.3) gives

$$\frac{d^2V}{dx^2}=-\frac{q}{\varepsilon\varepsilon_0}\left(N_d-2n_i \operatorname{Sinh}\frac{qV}{kT}\right) \qquad \cdots \quad (4.1.5)$$

This is the space charge equation describing the relation between the potential and the doping density. Unfortunately, this general equation does not have a closed form solution but approximations may be made. Whenever space charge neutrality can be assumed equation (4.1.5) reduces to

$$\frac{N_d}{n_i}=e^{qV/kT}\left(1-\frac{e^{-qV/kT}}{e^{qV/kT}}\right)=e^{qV/kT}\left(1-\frac{\bar{p}}{\bar{n}}\right) \qquad \cdots \quad (4.1.6)$$

On the other hand, point to point space charge neutrality implies everywhere

$$N_d+\bar{p}=\bar{n} \qquad \cdots \quad (4.1.7)$$

that is, point to point total positive and total negative charge equality. Equilibrium situation demands

$$\bar{p}=\frac{n_i^2}{\bar{n}} \qquad \cdots \quad (4.1.8)$$

Combining (4.1.7) and (4.1.8) we have

$$\bar{n}^2-\bar{n}N_d-n_i^2=0 \qquad \cdots \quad (4.1.9)$$

which gives ignoring the negative value for $\bar{n}$

$$\bar{n}=\frac{N_d}{2}\left[1+\sqrt{1+\frac{4(n_i^2)}{N_d}}\right] \qquad \cdots \quad (4.1.10)$$

This equation shows that while N_d is position dependent so long $N_d \gg n_i$, $\bar{n} \simeq N_d$. Say, in the worst case $N_d=10\ n_i$ at a position, then the electron concentration from (4.1.10) is $\bar{n}=1.01\ N_d$ while

$$\bar{p}=\frac{n_i^2}{\bar{n}}=\frac{n_i^2}{1.01\ N_d}$$

Even for a worst case position in the diode, we have from (4.1.3) or (4.1.6)

$$e^{qV/kT}=1.01\ N_d/n_i$$

$$\text{or} \quad \frac{qV}{kT}=\log(1.01)+\log\frac{N_d}{n_i}\simeq\log\frac{N_d}{n_i} \qquad \cdots \quad (4.1.11)$$

showing that point-to-point space charge neutrality requirement establishes a simple relationship between potential and doping density. Equation (4.1.11) shows that if the doping is varying exponentially, say, as

$$N_d = N_{de}\, e^{-2\theta x/W} \qquad \cdots \quad (4.1.12)$$

then V would be linearly varying with x and so we shall have a constant built in field. This constant field in turn will give $\frac{dE}{dx} = 0$ *i.e.*, point-to-point zero space charge. The value of the built in field for exponential doping can be obtained from (4.1.1) and (4.1.12) as

$$E = \frac{2\theta}{W} \cdot \frac{kT}{q} \qquad \cdots \quad (4.1.13)$$

also $\theta = 0$ implies uniform doping and zero built in field.

Equation (4.1.5) reminds us of the boundary. Due to the gradient of majority carriers, there is diffusion of these carriers at the boundaries. Equilibrium is attained by exposure of the positive space charge around $x = 0$ and accumulation of negative charge to the right of $x = W$ as shown in Fig. (4.1.1). It is these space charges where the field lines originate from and terminate to respectively.

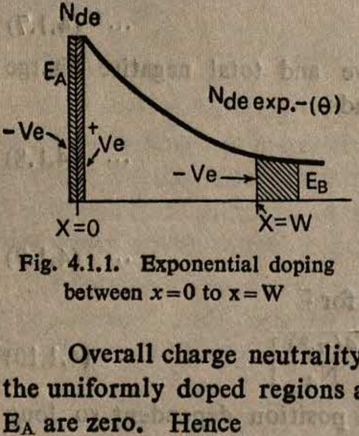


Fig. 4.1.1. Exponential doping between $x = 0$ to $x = W$

Overall charge neutrality is also maintained since the field on the uniformly doped regions at the two ends of the sample E_B and E_A are zero. Hence

$$\int_A^B \frac{dE}{dx}\, dx = E_B - E_A = 0 \qquad \cdots \quad (4.1.14)$$

4.2 Generalised P-N diode

If we have a p-n junction followed by a graded doped body (base) between two metal electrodes giving ohmic contact to the material as shown in Fig. 4.2.1, then from the ideas developed in the previous section, the space charge, potential and field variation along the length of diode can be idealized as shown in this diagram. The guiding lines are

(1) that there is a p-n junction formed having negative space charge in P-type material and positive in N-type material in the depletion region ;

(2) that the potential rises from P-type to N-type in this region ;

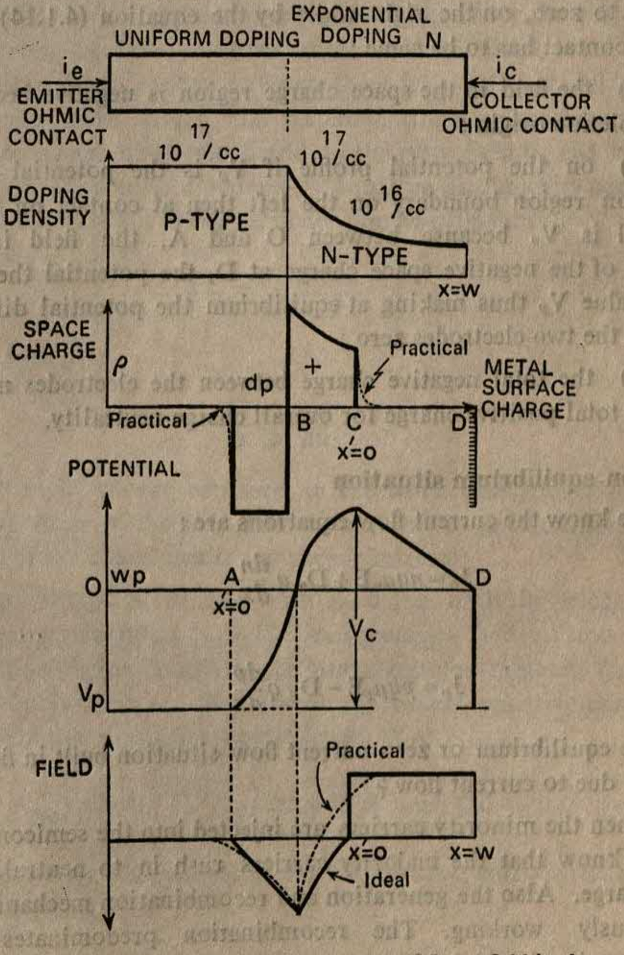


Fig. 4.2.1. Doping level and space charge, potential and field in the transition region of an exponentially doped *p-n* diode.

(3) because of exponential doping, the potential in N-type neutral region varies linearly with distance almost upto ohmic contact at the right ;

(4) since the field is positive and finite at this contact the field lines would require some negative charge for their termination while these lines would get numerous electrons on the metal surface and hence this would be a surface charge ;

(5) the field at the metal to semiconductor contact is thus reduced to zero, on the right while by the equation (4.1.14) that at the left contact has to be same ;

(6) the field in the space charge region is negative because it is of depletion origin ;

(7) on the potential profile if V_p is the potential at the transition region boundary on the left then at contact on left the potential is V_p because between O and A, the field is zero. Because of the negative space charge at D, the potential there falls to the value V_p thus making at equilibrium the potential difference between the two electrodes zero ;

(8) the total negative charge between the electrodes must be equal to total positive charge for overall charge neutrality.

4.3 Non-equilibrium situation

We know the current flow equations are :

$$J_n = nq\mu_n E + D_n q \frac{dn}{dx}$$

and ··· (4.3.1)

$$J_p = pq\mu_p E - D_p q \frac{dp}{dx}$$

Does the equilibrium or zero current flow situation built in field get modified due to current flow ?

When the minority carriers are injected into the semiconductor body we know that the majority carriers rush in to neutralise the space-charge. Also the generation and recombination mechanism are continuously working. The recombination predominates over generation in forward bias (equation 3.5.11) so the number of majority carriers lost at each moment needs to be replenished as demanded by the very small extra electric field created by the steady concentration of excess minority carriers at a place because there is certain injection (minority) level maintained at the boundary. It is to be remembered that if there were no recombination then such an extra field would have been non-existent. The worst case is, when carriers are moving slowest, that is when there is no built in field and E in (4.3.1) is small, the carriers are moving due to diffusion mainly. Assuming negligible recombination (implies use of very good single crystals). A close approximation for the value of the field E in equation (4.3.1) may be obtained by equating the majority current to zero because in such case the replenishing majority carriers do not need to flow. When there is a built in field, the carriers would move faster, the recombination replenishment of majority carriers are smaller still and so the extra field is, and hence the resultant field is again obtained by equating the majority current to zero. In our case for N-region under non-equilibrium

$$E = -\frac{kT}{q}\frac{1}{n}\frac{dn}{dx} \qquad \cdots \quad (4.3.2)$$

while for P-type bulk

$$E = \frac{kT}{q}\frac{1}{p}\frac{dp}{dx} \qquad \cdots \quad (4.3.3)$$

For example, we can estimate, in the worst case situation the order of magnitude of the field and the current, to get an insight in the meaning of the above mentioned approximation.

Say, in N-type Si sample with $\theta = 0$ (no built in field), the uniform doping density is $N_d = 10^{15}$/cc (hence $\bar{p} \sim 10^5$/cc) and say, at the transition region boundary the excess injected minority (hole) carrier density is 10^{14}/cc. We assume the excess minority carrier density distribution along the sample length W as

$\widehat{p} = 10^{14} f\left(1 - \frac{x}{W}\right)$ where $f\left(1 - \frac{x}{W}\right)$ is unity at $x = 0$ and zero at $x = W$

while for neutrality the majority carrier density will increase by equal amount giving

$$n = \bar{n} + \hat{n} = 10^{15} + 10^{14} f\left(1 - \frac{x}{W}\right) \text{ hence } p = 10^5 + 10^{14} f\left(1 - \frac{x}{W}\right)$$

Equating $J_n = 0$ (no recombination) we have

$$E = -\frac{kT}{q} \cdot \frac{1}{n} \cdot \frac{dn}{dx} = \frac{kT}{qW} \cdot \frac{f'}{(10+f)}$$

where $f' = -W\frac{\delta f}{\delta x}$.

This field gives the following values for the four current components

$$J_{p(\text{drift})} = q\mu_p(\bar{p} + \hat{p})E = q\frac{D_p(\bar{p} + 10^{14}f)f'}{W(10+f)}$$

$J_{p(\text{diff.})} = \frac{qD_p}{W}\,10^{14}f'$, with $\bar{p} = 10^5$/cc while it may be neglected at the injection boundary.

$$J_{n(\text{drift})} = q\mu_n(\bar{n} + \hat{n})E = q\frac{D_n}{W}\,10^{14}f'$$

$$J_{n(\text{diff.})} = -q\frac{D_n}{W}\,10^{14}f'$$

Where the value of f' at $x = 0$ and $x = W$ are f'_0 and f'_w respectively, we can write the current components as

at $x = 0$	and at $x = W$
$J_{p(\text{drift})} = \frac{qD_p}{W} f_0' \times 9{\cdot}1 \times 10^{12}$	$J_{p(\text{drift})} = \frac{qD_p}{W} f_w' \times 10^4$
$J_{p(\text{diff.})} = \frac{qD_p}{W} f_0' \times 10^{14}$	$J_{p\,(\text{diff}).} = \frac{qD_p}{W} f_w' \times 10^{14}$
$J_{n(\text{drift})} = \frac{qD_n f_0'}{W} \times 10^{14}$	$J_{n(\text{drift})} = \frac{qD_p}{W} \cdot f_w' \times 10^{14}$
$= -J_{n\,(\text{diff.})}$	$= -J_{n(\text{diff})}$.

Thus we see that the total current is composed mainly of diffusion current of minority carriers since the minority drift current is at least one order of magnitude lower than this. The field is very small (check with f, *a* linear distribution, *i.e.*, $f' = 1$ and $W = 10^{-3}$ cm) when there is no recombination and this is sufficient to cause two large equal and opposite components of majority currents to flow. If there is recombination to the extent of even 10% of the total current then the recombination part replenishment is possible by only less than 1% change in the existing field. Thus we see that estimation of the field by equation (4.3.2 or 4.3.3) is quite good even in the worst case ($\theta = 0$).

The other approximations that we shall use are that, (2) the boundary shift of the transition region is small enough to be negligible, (3) the conductivity of the doped regions are high enough not to cause appreciable voltage drops hence all of the applied voltage drops in the junction region because the conductivity there, is low for depletion of carriers, (4) the surface charge on the right-hand ohmic contact is negative that is of electrons in N-type semiconductor hence gives increased conductivity, while holes reaching there from the left find a sink. (5) the contacts are ohmic and acts as sinks for carriers.

4.4 Current Flow Equations

We have seen that when the minority carriers are injected in the N-type semiconductor, the minority hole carriers are the sole contributors to the total current only when the recombination is nil. Similarly, in the P-type body the minority (electron) carriers are the sole contributors under no recombination. Therefore, the minority current in an N-type sample on minority injection is

$$J_p = q\mu_p E(\bar{p} + \hat{p}) - D_p q \frac{d(\bar{p} + \hat{p})}{dx}$$

but equation (4.1.1) can be written as

$$q\mu_p E\bar{p} - D_p q . \frac{d\bar{p}}{dx} = 0.$$

For constant E by subtraction the expression for the current under non-equilibrium can, therefore, be written as

$$Jp = \hat{p} q\mu_p E - D_p q \frac{d\hat{p}}{dx} \quad \cdots \quad (4.4.1)$$

where $\hat{p}$ stands for excess minority carriers, d.c. or a.c. depends on the context. We have with exponential doping, a constant built in field

$$E = -\frac{kT}{q} \frac{1}{N_a} . \frac{dN_a}{dx} = \frac{2\theta}{W} . \frac{kT}{q} \quad \cdots \quad (4.1.13)$$

where $\theta = 0$ implies uniform doping.

Using eqn. (4.1.13), equation (4.4.1) can be written in the form

$$J_p = -qD_p \left(\frac{\hat{p}}{N_a} . \frac{dN_a}{dx} + \frac{d\hat{p}}{dx} \right) \quad \cdots \quad (4.4.2)$$

$$\text{and} \quad J_p = -qD_p\left(\frac{d\hat{p}}{dx} - \frac{2\theta}{W}\hat{p}\right) \qquad \cdots \quad (4.4.3)$$

If there is recombination then the electron current responsible for replenishment of the lost electrons has to be taken into account to get the total current.

Similarly, for the P-type semiconductor body, the built in field

$$E = \frac{kT}{q} \cdot \frac{1}{N_a} \cdot \frac{dN_a}{dx} = \frac{2\theta'}{W_p} \cdot \frac{kT}{q} \qquad \cdots \quad (4.4.4)$$

where θ' and W_p are the corresponding field factor and base length of the P-type material.

While the zero recombination current equations on minority injection are :

$$J_n = qD_n\left(\frac{d\hat{n}}{dx} + \frac{\hat{n}}{N_a}\frac{dN_a}{dx}\right) \qquad \cdots \quad (4.4.5)$$

$$\text{and} \quad J_n = qD_n\left(\frac{d\hat{n}}{dx} + \frac{2\theta'}{W_p} \cdot \hat{n}\right) \qquad \cdots \quad (4.4.6)$$

4.5 Transmission line Analogy

It can be shown that the minority carrier transport through the graded semiconductor body is similar to current flow through a tapered RCG transmission line. In such a transmission line if R_x and C_x are the resistance and capacitance per unit length respectively at x, and v the voltage across the line as shown in Fig. (4.5.1), the voltage change in an infinitesimal small length dx is $dv = -iR_x dx$

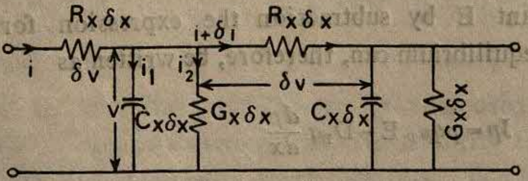


Fig. 4.5.1. Transmission line representation of an elemental length of the semiconductor

$$\text{or} \quad i = -\frac{1}{R_x} \cdot \frac{dv}{dx} = \frac{\text{Cm}}{\text{Ohms}} \cdot \frac{\text{Volts}}{\text{Cm}} = \text{Amps} \qquad \cdots \quad (4.5.1)$$

$$v=\frac{Q_x dx}{C_x dx}=\frac{\text{Coulomb}}{\text{Farad}}=\text{volt} \qquad \cdots \quad (4.5.2)$$

Substituting for v from (4.5.2) in equation (4.5.1) we have

$$i=\frac{1}{R_x C_x}\left(\frac{Q_x}{C_x}\frac{dC_x}{dx}-\frac{dQ_x}{dx}\right) \qquad \cdots \quad (4.5.3)$$

Write equation (4.4.2) as

$$i_p=D_p\left[-\frac{qA\widehat{p}}{N_a}.\frac{dN_a}{dx}-\frac{d}{dx}(qA\widehat{p})\right] \qquad \cdots \quad (4.5.4)$$

Substituting by inspection of (4.5.3) and (4.5.4)

$$\left.\begin{aligned} &D_p=\frac{1}{R_x C_x}\ \text{Cm}^2/\text{sec.}\\ &\frac{b}{N_a}=C_x\cdot\frac{\text{Farad}}{\text{Cm}} \ \text{ or } \ \frac{1}{N_a}\cdot\frac{dN_a}{dx}=-\frac{1}{C_x}\cdot\frac{dC_x}{dx}\\ &\text{and } \ qA\widehat{p}=Q_x\frac{\text{Coulomb}}{\text{Cm}} \end{aligned}\right\} \qquad \cdots \quad (4.5.5)$$

in Equation (4.5.4) we have

$$i_p=\frac{1}{R_x C_x}\left[\frac{Q_x}{C_x}.\frac{dC_x}{dx}-\frac{dQ_x}{dx}\right] \qquad \cdots \quad (4.5.6)$$

See that the equation (4.5.6) is the same as equation (4.5.3) showing that the per unit length resistance follows the impurity concentration, and capacitance inversely follows the impurity concentration. It is meant, therefore, that zero built in field implies uniform doping and hence analogous with the uniform line.

The diffusion term in the current flow equation may originate either from recombination or due to the presence of a sink or due to both these causes. But in our current flow equation (4.5.4), the recombination part does not appear explicitly, therefore the leakage conductance G_x per unit length, which stands for recombination, did not enter the equation (4.5.6) or (4.5.3) in explicit form.

To examine this point further we set up the time dependent current equation for the transmission line as follows.

From Fig. 4.5.1 we have

$$\left.\begin{aligned} &i_1=\frac{\delta}{\delta t}(C_x\,\delta x.v)=C_x\delta x\frac{\delta v}{\delta t}\\ \text{and} \quad &i_2=G_x\,\delta x v \end{aligned}\right\} \qquad \cdots \quad (4.5.7)$$

Using Kirchoff's law we have from Fig. 4.5.1

$$\left.\begin{aligned} &(i+\delta i)-i+i_1+i_2=0 \\ \text{hence} \quad &-\delta i=i_1+i_2 \end{aligned}\right\} \quad \ldots \quad (4.5.8)$$

Substituting (4.5.7) in equation (4.5.8) and dividing by δx we have

$$-\frac{\delta i}{\delta x}=C_x\frac{\delta v}{\delta t}+G_x v \quad \ldots \quad (4.5.9)$$

Using equation (4.5.2) for v in (4.5.9) and rearranging terms, we have

$$\frac{\delta Q_x}{\delta t}=-\frac{\delta i}{\delta x}-\frac{Q_x}{(C_x/G_x)} \quad \ldots \quad (4.5.10)$$

Substituting for $Q_x=qA\hat{p}$ and $T_p=C_x/G_x$ in (4.5.10) and dividing by (Aq) throughout we have

$$\frac{\delta\hat{p}}{\delta t}=-\frac{\delta J}{q\delta x}-\frac{\hat{p}}{T_p} \quad \ldots \quad (4.5.11)$$

This equation can be compared with equation (3.5.12) and seen that the transmission line continuity equation gives the semiconductor medium current continuity equation. The above considerations thus undoubtedly establish the identity of the transmission line and minority injected semiconductor body meaning thereby that the semiconductor body is entirely passive medium.

4.6 Minority carrier distribution

The recombination devoid model would give us a carrier distribution along the sample length as shown below. The current is entirely minority carrier current and from equation (4.4.2)

$$J_p=-D_p q\left(\frac{d\hat{p}}{dx}+\frac{\hat{p}}{N_d}\cdot\frac{dN_d}{dx}\right) \quad \ldots \quad (4.6.1)$$

which can be integrated, regarding J_p as constant in space, *i.e.*, $\frac{dJ_p}{dx}=0$ as needed for this model as

$$\int J_p N_d\,dx=-D_p q\int d(\hat{p}N_d)$$

$$\text{or} \quad \hat{p}=-\frac{J_p}{D_p q}\,\frac{\int^x N_d\,dx}{N_d}+\frac{C}{N_d} \quad \ldots \quad (4.6.2)$$

where the integration constant C is to be evaluated from the minority concentration value at the boundary at $x = \mathrm{W}$ in Fig. (4.2.1), which is the metal semiconductor ohmic contact sink, hence $\hat{p} = 0$.

Thus C becomes $$\mathrm{C} = \frac{\mathrm{J}_p}{\mathrm{D}_p q \mathrm{N}_a} \int^{\mathrm{W}} \mathrm{N}_a \, dx \qquad \cdots \quad (4.6.3)$$

Substituting (4.6.3) in equation (4.6.2)

$$\hat{p} = \frac{\mathrm{J}_p}{\mathrm{D}_p q} \frac{\int_x^{\mathrm{W}} \mathrm{N}_a \, dx}{\mathrm{N}_a} \qquad \cdots \quad (4.6.4)$$

Note that to find $\hat{p}$ and x, integrate from W to x,

at $x = 0$, $\hat{p} = \hat{p}_0$ then we have

$$\hat{p}_0 = \frac{\mathrm{J}_p}{\mathrm{D}_p q \mathrm{N}_{ae}} \int_0^{\mathrm{W}} \mathrm{N}_a \, dx \qquad \cdots \quad (4.6.5)$$

From (4.6.4) & (4.6.5) we have

$$\hat{p} = \hat{p}_0 \left(\frac{\mathrm{N}_{ae}}{\mathrm{N}_a}\right) \frac{\int_x^{\mathrm{W}} \mathrm{N}_a \, dx}{\int_0^{\mathrm{W}} \mathrm{N}_a \, dx} \qquad \cdots \quad (4.6.6)$$

For our assumed simple exponential doping $\mathrm{N}_a = \mathrm{N}_{ae} \left(e^{\frac{2\theta x}{\mathrm{W}}}\right)$ we have on integration and re-arrangement

$$\hat{p} = \hat{p}_0 \, e^{\frac{\theta x}{\mathrm{W}}} \frac{\sinh \theta\left(1 - \frac{x}{\mathrm{W}}\right)}{\sinh \theta} \qquad \cdots \quad (4.6.7)$$

expressed in terms of current from eqn. (4.6.5)

$$\hat{p} = \frac{\mathrm{J}_p . \mathrm{W}}{\mathrm{D}_p q} \cdot \frac{e^{-\left(1 - \frac{x}{\mathrm{W}}\right)}}{\theta} \sinh\left[\theta\left(1 - \frac{x}{\mathrm{W}}\right)\right] \qquad \cdots \quad (4.6.8)$$

In the recombination devoid model J_p is the total current and is constant along the sample length. Equation (4.6.8) enables one to visualise the excess minority concentration variation along the sample length for different values of θ, the built in field as a parameter, on the assumptions that in all cases the same current is flowing. This is shown in Fig. (4.6.1).

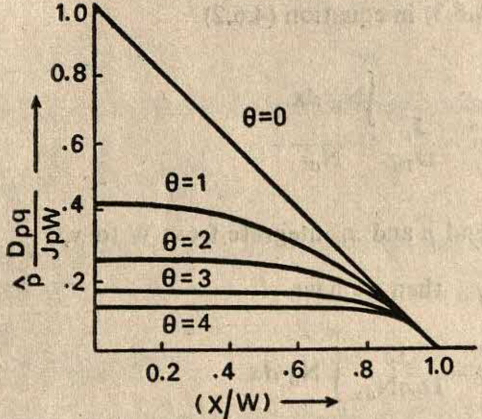


Fig. 4.6.1. Concentration variation of minority carriers in no recombination case.

Note that in zero recombination model, the diffusion limited flow ($\theta = 0$) gives a linear fall of excess minority carrier concentration. As θ that is the built in field increases the gradient of concentration decreases in the semiconductor body, particularly very fast, near the injection end. At $\theta = 0$ for over 80% of the body the concentration is nearly constant meaning that over that region current flow is due mainly to drift of carriers, while for all values of θ the excess carrier distributions coincide with that for the diffusion limited case near the collection end. This is because the excess concentration there, is tending towards zero hence the field current, being proportional to the concentration, is also becoming nil yielding place to diffusion entirely.

It is also seen that increasing the built in field above that for $\theta = 4$ does not appreciably increase the drift region in the diode. Hence in practice $\theta = 4$, is the highest value for θ one uses.

4.7 D.C. current voltage relation

We can now find the d.c. current voltage relationship for *p-n* junction diode. For the current due to injection of minority carriers in *n*-type body we write from equation (4.6.5)

$$J_p = \frac{D_p q N_{de} \hat{p}_o}{\int_0^W N_d \, dx} \quad \cdots \quad (4.7.1)$$

The excess hole concentration $\hat{p}_o$ at the boundary at $x=0$ in Fig. (4.2.1) from equation (3.3.11)

$$\hat{p}_o = \bar{p}_{no}(e^{qV_e/kT} - 1) \quad \cdots \quad (4.7.2)$$

From equations (4.7.1) and (4.7.2) setting $\bar{p}_{ro} . N_{de} = n_i^2$

$$J_p = \frac{D_p q n_i^2 (e^{qV_e/kT} - 1)}{\int_0^W N_d \, dx} \quad \cdots \quad (4.7.3)$$

Similarly, for the current due to simultaneous injection of electrons in the *p*-type body, on the left in Fig. (4.2.1), can be written like (4.7.3)

$$J_n = \frac{D_n q n_i^2 (e^{qV_e/kT} - 1)}{\int_0^{W_p} N_a \, dx} \quad \cdots \quad (4.7.4)$$

The minority holes in *n*-type body are flowing from the junction boundary $x=0$ to the right-hand ohmic contact $x=W$ while the minority electrons in *p*-type are moving from the depletion boundary on the left $x'=0$ at A to the left-hand ohmic contact at $x=W_p$. Hence the two current contributions are additive Hence the total current in *p-n* junction diode can be written, by multiplying by collection area A as

$$I = I_n + I_p = q n_i^2 A (e^{qV_e/kT} - 1) \left[\frac{D_p}{\int_0^W N_d \, dx} + \frac{D_n}{\int_0^{W_p} N_a \, dx} \right] \quad \cdots \quad (4.7.5)$$

with $$I_o = Aqn_i^2\left[\frac{D_p}{\int_0^W N_d\,dx} + \frac{D_n}{\int_0^{W_p} N_a\,dx}\right] \qquad \cdots \quad (4.7.6)$$

Equation (4.7.5) can be written as

$$I = I_0(e^{qV_e/kT} - 1) \qquad \cdots \quad (4.7.7)$$

4.8 Injection Efficiency

Our main object is to inject a type of carrier and their control hence we define the injection efficiency as the ratio of injected useful current to the total current as

$$Y = \frac{I_p}{I} = \frac{1}{1 + \frac{D_n}{D_p}\left[\int_0^W N_d\,dx / \int_0^{W_p} N_a\,dx\right]} \qquad \cdots \quad (4.8.1)$$

From equation (4.8.1) we see that by making the total number of acceptor atoms on P-side very much larger than the total number of donor atoms on N-side the injection efficiency can be improved to near unity value. In that case the diode current will be mainly minority current holes flowing in the *n*-type base bulk.

4.9 Solution of continuity equation

We need to examine the effect of recombination as also the effect of variation of the injection level at the boundary. So, we should take a time dependent equation containing the recombination terms. The continuity equation (3.5.12) for flow of holes in the *n*-type base reads,

$$\frac{dp}{dt} = -\frac{1}{q}\frac{dJ_p}{dx} - \frac{\hat{p}}{T_p} \qquad \cdots \quad (4.9.1)$$

From equation (4.4.3) we have

$$J_p = D_p q\left[\frac{d\hat{p}}{dx} - \frac{2\theta\hat{p}}{W}\right] \qquad \cdots \quad (4.9.2)$$

Substituting equation (4.9.2) in (4.9.1) we have

$$\frac{dp}{dt} = D_p\left(\frac{d^2\hat{p}}{dx^2} - \frac{2\theta}{W}\frac{d\hat{p}}{dx}\right) - \frac{\hat{p}}{T_p} \qquad \cdots \quad (4.9.3)$$

For electrons in a p-type medium a similar equation can be found from equation (3.5.13) and (4.4.6) as

$$\frac{d\widehat{n}}{dt} = -D_n\left(\frac{d^2\widehat{n}}{dx^2} + \frac{2\theta'}{W_p}\frac{d\widehat{n}}{dx}\right) - \frac{\widehat{n}}{T_n} \qquad \cdots \quad (4.9.4)$$

For an n-type medium we solve for equation (4.9.3). The excess concentration at any point can be steady (d.c.) or fluctuating (a.c. superimposed on d.c.). We, therefore, write

$$\widehat{p} = p_{dc} + p_{ac} = p_{dc} + \overleftharpoon{p}\, e^{j\omega t} \qquad \cdots \quad (4.9.5)$$

For the d.c. case, the equation that applies will be obtained from equation (4.9.3) as

$$0 = D_p\left(\frac{d^2 p_{dc}}{dx^2} - \frac{2\theta}{W}\frac{dp_{dc}}{dx}\right) - \frac{p_{dc}}{T_p} \qquad \cdots \quad (4.9.6)$$

For the more general case we have from equation (4.9.3) on substitution of $\widehat{p}$ from equation (4.9.5)

$$\frac{dp_{ac}}{dt} = D_p\left[\frac{d^2 p_{dc}}{dx^2} + \frac{d^2 p_{ac}}{dx^2} - \frac{2\theta}{W}\frac{d}{dx}(p_{dc} + p_{ac})\right]$$
$$- \frac{p_{dc}}{T_p} - \frac{p_{ac}}{T_p} \qquad \cdots \quad (4.9.7)$$

In this equation the terms containing p_{dc} satisfy equation (4.9.6), that is, they equate to zero, leaving the a.c. equation only, *i.e.*

$$\frac{dp_{ac}}{dt} = D_p\left[\frac{d^2 p_{ac}}{dx^2} - \frac{2\theta}{W}\frac{dp_{ac}}{dx}\right] - \frac{p_{ac}}{T_p} \qquad \cdots \quad (4.9.8)$$

which with $p_{ac} = \overleftharpoon{p}\, e^{j\omega t}$ reduces to

$$j\omega\overleftharpoon{p} = D_p\left[\frac{d^2\overleftharpoon{p}}{dx^2} - \frac{2\theta}{W}\frac{d\overleftharpoon{p}}{dx}\right] - \frac{\overleftharpoon{p}}{T_p} \qquad \cdots \quad (4.9.9)$$

Note that equation (4.9.8) can be solved with a.c. boundary conditions alone without any attention to the d.c. boundary values. Also note that the form of equation (4.9.9) reduces to that of equation (4.9.6) if in equation (4.9.9) ω is set to zero, therefore we write equation (4.9.9) as

$$j\omega P = D_p\left[\frac{d^2 P}{dx^2} - \frac{2\theta}{W}\frac{dP}{dx}\right] - \frac{P}{T_p} \qquad \cdots \quad (4.9.10)$$

and bear in mind that after solving for P the solution can be used for both d.c. and a.c. In the former set $\omega=0$ and $P=p_{dc}$ while in the latter set $P=\overleftharpoonup{p}$.

Equation (4.9.9) can be rewritten using $D_p.T_p=L_p^2$ as

$$P\frac{(1+j\omega T_p)}{L_p^2}=\frac{d^2P}{dx^2}-\frac{2\theta}{W}\frac{dP}{dx} \qquad \cdots \quad (4.9.11)$$

Substituting for $\left(\frac{1+j\omega T_p}{L_p^2}\right)$ with a^2 $\qquad \cdots \quad (4.9.12)$

equation (4.9.11) becomes

$$\frac{d^2P}{dx^2}-\frac{2\theta}{W}\frac{dP}{dx}-a^2P=0 \qquad \cdots \quad (4.9.13)$$

For the d.c. case $a^2=1/L_p^2$, while for the a.c. case

$$a^2=(1+j\omega T_p)/L_p^2$$

Using the solution of the form e^{mx} and solving for the roots, the solution of equation (4.9.13) may be written as

$$P=e^{\frac{\theta}{W}x}\left(Ae^{\frac{\phi}{W}x}+Be^{-\frac{\phi}{W}x}\right) \qquad \cdots \quad (4.9.14)$$

$$\text{where} \quad \psi=(\theta^2+a^2W^2)$$

Next we need to evaluate the constants in equation (4.9.14) from the boundary concentrations. At $x=W$, both p_{dc} or $\overleftharpoonup{p}$ are zero hence $P=0$ while at $x=0$, say, $P=P_o$. For a.c. case solution we will use

$$\cdots \qquad P_0=\overleftharpoonup{p}_0=\frac{q}{kT}\hat{p}_0\,\overleftharpoonup{v}$$

obtained from equation (3.3.9) with $\hat{p}_0=\bar{p}_{no}\left(\exp\frac{qV_e}{kT}-1\right)$ from equation (3.3.10). For d.c. case solution substitute

$P_o=\hat{p}_o=\bar{p}_{no}\,[\exp\,(qVe/kT)-1)]$ from equation (3.3.10) with the aid of these boundary conditions equation (4.9.14) can be expressed as

$$P=P_o\,e^{\frac{\theta x}{W}}\left[\frac{\sinh\varphi(1-x/W)}{\sinh\varphi}\right] \qquad \cdots \quad (4.9.15)$$

From this equation the d.c. solution can be written as

$$p_{dc}=\hat{p}=\bar{p}_{no}[\exp(qVe/kT-1)e^{\frac{\theta x}{W}}\left[\frac{\sinh\varphi_o(1-x/W)}{\sinh\varphi_o}\right] \qquad \cdots \quad (4.9.16)$$

$$\text{where} \quad \varphi_o{}^2=(\theta^2+W^2/L_p{}^2) \qquad \cdots \quad (4.9.17)$$

Similarly the a.c. solution can be written as

$$\tilde{p}=\frac{\tilde{v}_o q\bar{p}_n}{kT}[\exp(qV_e/kT)-1)]e^{\theta x/W}\left[\frac{\sinh\psi(1-x/W)}{\sinh\psi}\right] \quad \cdots \quad (4.9.18)$$

where $\psi^2=[\theta^2+(1+j\omega T_p)W^2/L_p{}^2$ $\cdots$ (4.9.19a)

$$=\psi_o{}^2\left[1+j\omega\frac{W^2/D_p}{\psi_o{}^2}\right] \quad \cdots \quad (4.9.19b)$$

4.10 General expression for Current

We have obtained the general expression for the excess concentration variation along the sample length as

$$P=P_o e^{\theta x/W}\left[\frac{\sinh\psi(1-x/W)}{\sinh\psi}\right] \quad \cdots \quad (4.9.15)$$

and we have minority hole current as

$$I_p=D_p qA\left[\frac{dP}{dx}-\frac{2\theta}{W}P\right] \quad \cdots \quad (4.10.1)$$

Substituting the expression for P from (4.9.15) and simplifying we have,

$$I_p=P_o\frac{D_p qA}{W}e^{\frac{\theta}{W}}\left[\frac{\theta\sinh\psi(1-x/W)+\psi\cosh\psi(1-x/W)}{\sinh\psi}\right] \quad \cdots \quad (4.10.2)$$

This is the general expression for current. It is a function of position because there is recombination and so

$$\frac{dJ_p}{dx}\neq 0$$

To find d.c. current at $x=0$, we put $\omega=0$ and $x=0$ in equation (4.10.2) with $P_o=\hat{p}_o=\bar{p}_{no}[\exp(qV_e/kT)-1]$ and hence

$$I_{pe}=\hat{p}_o\frac{D_p qA}{W}(\theta+\psi_o\coth\psi_o) \quad \cdots \quad (4.10.3)$$

To find d.c. current at $x=W$, we put $\omega=0$ and $x=W$ and $P_o=\hat{p}_o$ to obtain

$$I_{pc}=\frac{\hat{p}_o D_p qA}{W}\cdot\frac{\psi_o e^{\theta}}{\sinh\psi_o} \quad \cdots \quad (4.10.4)$$

The recombination replenishing d.c. electron current must be

$$I_n(x)=I_{pe}-I_p(x) \quad \cdots \quad (4.10.5)$$

For a.c. current at $x = 0$, put $x = 0$ and

$P_o = \tilde{p}_0 - \frac{q}{kT} \cdot \hat{p}_o \cdot \bar{v}$ in equation (4.10.2) to have,

$$i_{pe} = A\frac{q}{kT} \cdot \hat{p}_0 \frac{\tilde{v} D_p q}{W}(\theta + \psi \coth \psi) \qquad \cdots \quad (4.10.6)$$

Similarly for a.c. current at $x = W$, substitute $x = W$ and

$P_0 = \tilde{p}_0 - \frac{q}{kT} \hat{p}_0 \tilde{v}$ in eqn. (4.10.2) to obtain

$$i_{pc} = \frac{Aq}{kT} \cdot \hat{p}_0 \cdot \frac{\tilde{v} D_p q}{W} \cdot \frac{e^{\theta}\psi}{\sinh \psi} \qquad \cdots \quad (4.10.7)$$

The recombination replenishing a.c. electron current must be

$$i_n(x) = i_{pe} - i_p(x) \qquad \cdots \quad (4.10.8)$$

4.11 Transport factor

In general we can define the current transport factor as the ratio of minority current at collector to minority current at the emitter. Then the a.c. transport factor can be written from equation (4.10.6) and (4.10.7) as

$$\alpha = \frac{i_{pc}}{i_{pe}} = \frac{\psi e^{\theta}}{\theta \sinh \psi + \psi \cosh \psi} \qquad (4.11.1)$$

while for d.c. transport factor, we have from equation (4.10.3) and (4.10.4) or from (4.11.1) for $\omega = 0$

$$\alpha_0 = \frac{\psi_0 e^{\theta}}{\theta \sinh \psi_0 + \psi_0 \cosh \psi_0} \qquad \cdots \quad (4.11.2)$$

In the above equations

$$\psi^2 = [\theta^2 + (1 + j\omega T_p)W^2/L_p^2] = \left[\psi_0^2 + j\omega \frac{W^2}{D_p}\right]$$

and $\quad \psi_0^2 = [\theta^2 + W^2/L_p^2]$

and from eqn. (4.1.13) $\quad \theta = WqE/2kT \qquad \cdots \quad (4.11.3)$

which is a measure of built-in field.

Equation (4.11.2) indicates that α_0 is less than unity but so long $W/L_p \leqslant \theta$ which is a relatively easy condition to satisfy, $\alpha_0 \simeq 1$.

Thus nearly satisfying this condition one can have effective d.c. transport of minority carriers. The diffusion limited flow case *i.e.*, $\theta = 0$ case will be dealt in the next section.

4.12 Diffusion Limited Diode d.c. case

When the diode body is uniformly doped, the built-in field is zero ($\theta = 0$) and the minority carrier flow is mainly diffusion flow. In this case, the built-in field is zero, so that the carriers are not accelerated and hence the average motion of the carriers is slow providing more opportunity for recombination. This is why we refer to the diffusion flow case as the highest recombination case.

In this case, we can get the excess hole concentration variation along the *n*-type sample, from equation (4.9.15) with $\theta = 0$ and $P_o = \hat{p}_o$, for the d.c. situation as

$$\hat{p} = \hat{p}_o \frac{\sinh [W/L_p (1 - x/W)]}{\sinh (W/L_p)} \qquad \cdots \quad (4.12.1)$$

The nature of this variation, shown in Fig. (4.12.1a) can be easily understood if we examine the two extreme cases $W \gg L_p$ and $W \ll L_p$. In the first case, equation (4.12.1) simplifies to :

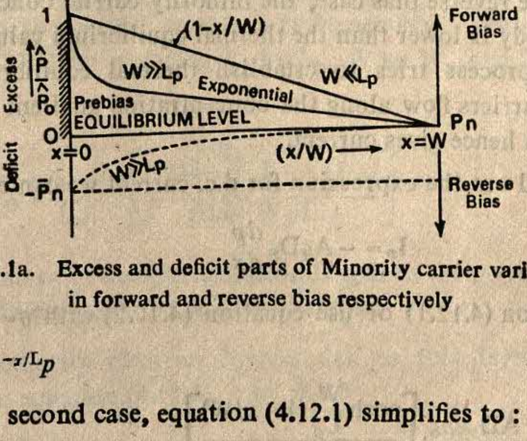


Fig. 4.12.1a. Excess and deficit parts of Minority carrier variation in forward and reverse bias respectively

$$\hat{p} = \hat{p}_o e^{-x/L_p} \qquad \cdots \quad (4.12.2)$$

while, in the second case, equation (4.12.1) simplifies to :

$$\hat{p} = \hat{p}_o (1 - x/W) \qquad \cdots \quad (4.12.3)$$

For $W \gg L_p$ we see that the concentration variation along the sample is exponential and justifies the definition of L_p, the diffusion length. In the (narrow base) $W \ll L_p$, we see that the concentration decreases linearly with distance. This diode is of more practical interest than the former because as we shall see in a later section, the transit time of minority carriers can be made small by decreasing W, the width of the base within practical limits.

For the reverse biased diode, the total concentration at the boundary tends towards zero and $\hat{p}_0$ becomes the deficit part of the carriers as shown in Fig. (4.12.1b).

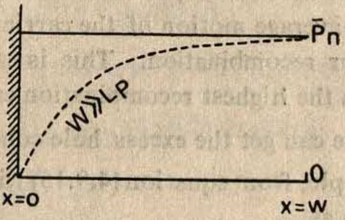


Fig. 4.12.1b. Minority carrier variation in reverse biased diode

The reverse current is called the generation current, as opposed to the forward current which is called the recombination current. This is true for reverse biased diode but for forward bias case it is strictly true for wide base diode only.

For the reverse bias case, the minority carrier concentration in the diode body is lower than the thermal equilibrium value, while the generation process tries to establish thermal equilibrium but the generated carriers flow along the concentration gradient towards the junction and hence gives current.

To find out the expression for d.c. current we can either use

$$I_p = -AqD_p \frac{dp}{dx} \qquad \cdots \quad (4.12.4)$$

with equation (4.12.1) or use equation (4.10.2) with $\theta = 0$ and $\omega = 0$ to have

$$I_p(x) = \frac{AD_p q\hat{p}_0}{L_p} \cdot \left[\frac{\cosh \frac{W}{L_p}(1 - x/W)}{\sinh W/L_p}\right] \qquad \cdots \quad (4.12.5)$$

The highest value of this occurs at $x = 0$ and is

$$I_{pe} = \frac{qD_pA}{L_p} \cdot \hat{p}_0 \coth W/L_p \qquad \cdots \quad (4.12.6)$$

Hence the majority electron current replenishing the recombination loss is

$$I_n(x) = I_{pe} - I_p(x) \qquad \cdots \quad (4.12.7)$$

It is zero at the point of injection and increases with distance by the amount, the hole current decreases as shown in Fig. (4.12.2).

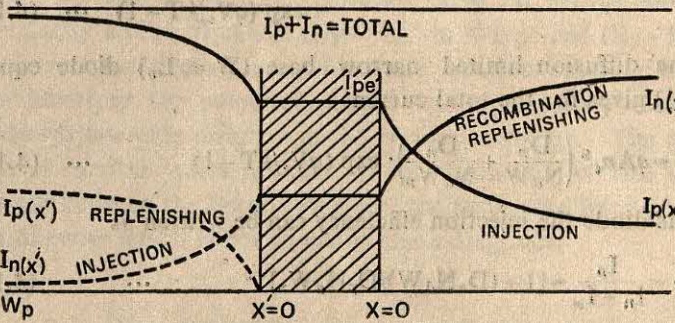


Fig. 4.12.2. Both injection current and recombination Replenishing current in the diode body

A similar equation can be written for minority electron current variation in the p-type sample simultaneously. With the aid of equation (4.12.5) by replacing L_p by L_n, W by W_p and D_p by D_n and $\hat{p}_0$ by $\hat{n}_0'$ as

$$I_n(x') = \frac{AD_n q \hat{n}_o}{L_n} \cdot \left[\frac{\cosh \frac{W_p}{L_n} (1 - x'/W_p)}{\sinh (W_p/L_n)} \right] \quad \cdots \quad (4.12.8)$$

where L_n = minority electron diffusion length in the p-type region and D_n = diffusion constant in the p-type region.

The minority electron current due to this injection can be written from equation (4.12.6). At $x' = 0$

$$I_{ne} = \frac{AD_n q}{L_n} \cdot \hat{n}_o . \coth W_p/L_n \quad \cdots \quad (4.12.9)$$

Adding the current contributions given by equations (4.12.6) and (4.12.9) and using

$$\left. \begin{aligned} \hat{p}_o &= \bar{p}_{no} \exp (qV_e/kT - 1) \\ \hat{n}_o &= \bar{n}_{po} \exp (qV_e/kT - 1) \\ \text{and} \quad \bar{p}_{no} N_d &= n_i^2 = \bar{n}_{po} N_a \end{aligned} \right\} \quad \cdots \quad (4.12.10)$$

we can write for the total current flowing as

$$I = I_{ps} + I_{ns} = qAn_i^2 \left(\frac{D_p}{L_p N_d} \cdot \coth \frac{W}{L_p} + \frac{D_n}{L_n N_a} \cdot \coth \frac{W_p}{L_n}\right) \times \exp (qV_e/kT - 1) \quad \cdots \quad (4.12.11)$$

For the diffusion limited narrow base ($W \leqslant L_p$) diode equation (4.12.11) gives for the total current

$$I = qAn_i^2 \left(\frac{D_p}{N_d W} + \frac{D_n}{N_a W_p}\right) \cdot \exp (qV_e/kT - 1) \quad \cdots \quad (4.12.12)$$

For this diode the injection efficiency can be written as

$$\gamma = \frac{I_p}{I_n + I_p} = [1 + (D_n N_d W)/D_p N_a W_p]^{-1} \quad \cdots \quad (4.12.13)$$

By making $N_a \geqslant N_d W/W_p'$ unity injection efficiency can be approached and a hole carrier diode may be obtained. Conversely by making $N_d \geqslant N_a W_p/W$ an electron carrier diode may be obtained. Thus we get a single type of current carrier diode.

The d.c. transport factor for diffusion limited diode can be written, from equation (4.11.2) with $\theta = 0$ as :

$$\alpha_0 = \frac{1}{\cosh W/L_p} \quad \cdots \quad (4.12.14)$$

which for the narrow base diode ($W \leqslant L_p$) can be expanded and written as

$$\alpha_0 = 1 - (W^2/2L_p^2) \quad \cdots \quad (4.12.15)$$

that is, in the narrow base diffusion limited diode, the transport factor improves rapidly as W decreases.

In this section it would also be worth considering the positions of the Fermi levels in forward and reverse biased wide base diodes. The variation of minority carriers in the wide base diode is exponentially related to distance as $\exp(-x/L_p)$ while their concentration is proportional to $\exp - (E_{FP} - E_V)/kT$, hence from the junction boundary into the body $(E_{FP} - E_V)$ should change linearly as shown in Figs. 4.12.3a and 4.12.3b. The majority carrier concentration being considered very high compared to the level of minority injection the majority concentration hence the distance $(E_C - E_{FN})$ does hardly change in the same medium. Note that, likewise in the forward biased case in the P-medium the distance between E_{FN} and E_C

increases as the carriers move away from the junction, linearly indicating exponential decrease of injected electrons while $(E_V - E_{FP})$ hardly changes in this medium. Similarly, in the reverse bias case since the minority density increases from about zero at the junction boundary to the thermal equilibrium value far within the body exponentially so the distances $(E_{FP} - E_V)$ in N-type and $(E_C - E_{FN})$ in P-type decreases linearly away from the junction boundaries. For abrupt junctions the potential changes as the difference of a square and a linear law with distance in the transition region. The concentration of carriers varies inversely as the exponential of this voltage so the gap between the Fermi level and the respective band edge also varies likewise which is not shown in these diagrams.

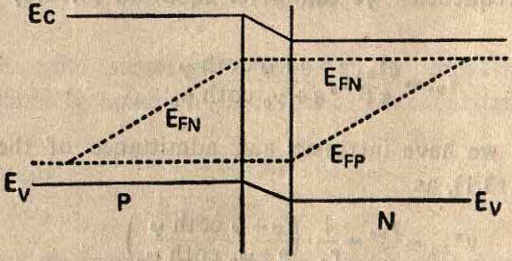


Fig. 4.12.3a. Variation of quasi fermi level (IMREF) in forward biased diode body

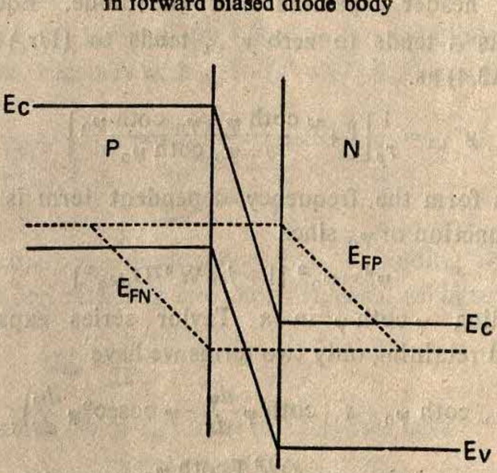


Fig. 4.12.3b. Variation of IMREF in reverse biased diode body

4.13 A.C. Impedance of diode

At very low frequencies the diode forward impedance is a resistance r_e and can be found from equation (4.7.7) as

$$\frac{1}{r_e}=\frac{\delta I}{\delta V_e}=\frac{q}{kT}(I+I_o)\simeq\frac{qI}{kT}$$

Assuming that the diode is a single carrier diode we have

$$\frac{I}{r_e}=\frac{q}{kT}\cdot I_{pe} \quad \cdots \quad (4.13.1)$$

At room temperature (300°K) for a d.c. current of $I_{pe} mA$

$$r_e=\frac{25}{I_{pe}(mA)}\cdot \text{Ohms} \quad \cdots \quad (4.13.2)$$

At higher frequencies we can write equation (4.10.6) with equation (4.10.3) as

$$i_{pe}=\frac{qI_{pe}}{kT}\tilde{v}_e\frac{\theta+\psi\coth\psi}{\theta+\psi_o\coth\psi_o} \quad \cdots \quad (4.13.3)$$

from which we have intrinsic a.c. admittance of the diode using equation (4.13.1), as

$$y^*_{11}=\frac{i_{pe}}{\tilde{v}_e}=\frac{1}{r_e}\cdot\left(\frac{\theta+\psi\coth\psi}{\theta+\psi_o\coth\psi_o}\right) \quad \cdots \quad (4.13.4)$$

Intrinsic, because this excludes the transition region capacitance and the metallic header capacitance of the diode. Equation (4.13.4) shows that as ω tends to zero y^*_{11} tends to $(1/r_e)$; so, we write equation (4.13.4) as

$$y^*_{11}=\frac{1}{r_e}\left[1+\frac{\psi\coth\psi-\psi_0\coth\psi_0}{\theta+\psi_0\coth\psi_0}\right] \quad \cdots \quad (4.13.5)$$

Even in this form the frequency dependent term is involved as a hyperbolic function of ψ_0 since

$$\psi^2=\psi_0^2\,[1+j\omega\, W^2/D_p\psi_0^2]$$

Expanding $\psi\coth\psi$ in a Taylor series expansion around $j\omega=s=0$ and retaining only two terms we have

$$\psi\coth\psi=\psi_0\coth\psi_0+s\left(\coth\psi\frac{d\psi}{ds}-\psi\,\text{cosec}^2\psi\frac{d\psi}{ds}\right)_{s=0} \quad (4.13.6)$$

$$\text{or}\quad(\psi\coth\psi-\psi_0\coth\psi_0)=\frac{sW^2}{2D_p}\left[\frac{\coth\psi_0}{\psi_0}-\text{cosech}^2\psi_0\right] \cdots \quad (4.13.7)$$

the region of convergence of coth ψ gives us the frequency

$$\omega < \frac{D_p}{W^2}(\psi_0{}^2+\pi^2) \qquad \cdots \quad (4.13.8)$$

as the usable range of frequency.

Substituting equation (4.13.7) in equation (4.13.5) with $s = j\omega$ we have

$$y^*_{11} = \frac{1}{r_e}\left[1 + j\omega\,\frac{W^2}{2D_p}\,\frac{\left(\dfrac{\coth\psi_0}{\psi_0} - \operatorname{cosech}^2\psi_0\right)}{\theta + \psi_0\coth\psi_0}\right] \qquad \cdots \quad (4.13.9)$$

Since Sinh $2\psi_0 \geqslant 2\psi_0$ we have $\cosh\psi_0 \geqslant \dfrac{\psi_0}{\sinh\psi_0}$

$$\text{or } \frac{\coth\psi_0}{\psi_0} \geqslant \operatorname{cosech}^2\psi_0 .$$

Therefore, the susceptance in y^*_{11} is capacitive for all frequencies defined by equation (4.13.8) and this capacitance is

$$C_{de} = \frac{W^2}{2D_p r_e}\left[\frac{\dfrac{\coth\psi_0}{\psi_0} - \operatorname{cosech}^2\psi_0}{\theta + \psi_0\coth\psi_0}\right] \qquad \cdots \quad (4.13.10)$$

In order to get an insight into the origin of this capacitance we examine in the following section the charge storage in the semiconductor body and its withdrawal.

As a matter of interest, for diffusion limited diode the value of this capacitance, becomes with $\psi_0{}^2 = (\theta^2 + W^2/L_p{}^2)$ and $\theta = 0$

$$C_{de} = \frac{W^2}{2D_p r_e}\left[\frac{\dfrac{L_p}{W}\coth\dfrac{W}{L_p} - \operatorname{cosech}^2\dfrac{W}{L_p}}{\dfrac{W}{L_p}\cdot\coth\dfrac{W}{L_p}}\right] \qquad \cdots \quad (4.13.11)$$

For narrow base diode ($W \leqslant L_p$), expanding the hyperbolic terms in series and retaining two terms in each, we have

$$C_{de} = \frac{W^2}{2D_p r_e}\left(\frac{2}{3}\right) = \frac{W^2}{3D_p r_e} \qquad \cdots \quad (4.13.12)$$

for wide base diode ($W \geqslant L_p$) from (4.13.11) we have

$$C_{de} = \frac{T}{2r_e}\left[1 - \frac{2W^2/L_p{}^2}{\sinh W/L_p}\right] \simeq \frac{T}{2r_e} \qquad \cdots \quad (4.13.13)$$

In equation (4.13.10) note that as ψ_0 tends to large values coth $\psi_0 \rightarrow$ unity write cosech^2 $\psi_0 \rightarrow$ zero, hence $C_{de} \rightarrow [W^2/2D_p r_e \psi_0 (\theta + \psi_0)]$ for $\frac{W}{L_p} \ll 1$ and $\psi_0 \simeq \theta_0 = 4$. We have $C_{de} = (W^2/64 D_p r_e) = \frac{W^2}{(2\theta)^2 D_p r_e}$ (c.f. equation 4.13.12) or in other words with increased built-in field the input capacitance decreases in value.

Note that the diode input admittance is a parallel combination of a resistance and a capacitance. The lower is the value of this capacitance, the better would be the frequency performance of the diode.

4.14 Stored and reclaimable charges

We have introduced minority charge carriers in the semiconductor body, during their transit both the neutralizing majority and introduced minority carriers were in the medium. We calculate the minority charge stored Q_S and expect to find a relationship of this charge with the transit time T_B. Assuming, both drift and diffusion results in an average space dependent velocity $\bar{v}$ we write

$$T_B = \int_0^W (dx/\bar{v}) \qquad \cdots \quad (4.14.1)$$

$$\text{while } I_{pe} = qA\ \hat{p}\ \bar{v} \qquad \cdots \quad (4.14.2)$$

From eqns. (4.14.1) and (4.14.2), we have

$$T_B = \frac{1}{I_{pe}} \int_0^W A\hat{p}\ q\ dx = \frac{Q_s}{I_{pe}} \qquad \cdots \quad (4.14.3)$$

$$\text{since the stored charge } Q_s = \int_0^W A\ \hat{p}\ q\ dx \qquad \cdots \quad (4.14.4)$$

Substituting the value for $\hat{p}$ from eqn. (4.9.16) and integrating (4.14.4), we have

$$Q_S = \frac{A\ \hat{p}_0\ q\ L_p^2}{W} [\theta + \psi_0 \coth \psi_0 - \psi_0\ e^{\theta} \operatorname{cosec} \psi_0] \qquad (4.14.5)$$

$$\text{while } I_{pe} = \frac{Aq\hat{p}_0 D_p}{W}\,[\theta + \psi_0 \coth \psi_0] \qquad \cdots \quad (4.14.6)$$

From eqn. (4.14.3) with eqn. (4.14.5), (4.10.3) and (4.11.2), we have

$$T_B = T_p\,(1 - \alpha_0) \qquad \cdots \quad (4.14.7)$$

Next we calculate the reclaimable minority charges that comes out from the two ends when the input voltage is stepped to zero from a steady value. For this, solve the continuity eqn. (4.9.3) for a step voltage down input to get the resultant hole concentration variation and its decay. For boundary conditions we know that before step input there was along the semiconductor body the excess concentration variation given by eqn. (4.9.16). After the application of the step the excess concentration at $x = 0$ and $x = W$ is zero for all times is assumed.

The solution comes out as

$$\hat{p} = 2\hat{p}_0 \sum_{1}^{\infty} \left[\frac{n\pi}{n^2\pi^2 + \psi_0^{\,2}}\, e^{\left\{ \frac{\theta}{W} x - (n^2\pi^2 + \psi_0^{\,2}) \frac{D_p t}{W^2} \right\}} \cdot \sin \frac{n\pi x}{W} \right] \qquad \cdots \quad (4.14.8)$$

The current transient i_{pe} at $x = 0$ the input end, is

$$i_{pe}\,(t) = -D_p\, qA \left(\frac{dp}{dx} \right)_{x=0} \qquad \cdots \quad (4.14.9)$$

since $\hat{p} = 0$ at $x = 0$ is being assumed with (4.14.8). Eqn. (4.14.9) gives on using eqn. (4.10.3)

$$i_{pe}\,(t) = \frac{-2I_{pe}}{(\theta + \psi_0 \coth \psi_0)} \sum_{1}^{\infty} \left[\frac{n^2\pi^2}{n^2\pi^2 + \psi_0^{\,2}}\, e^{\left\{ -(n^2\pi^2 + \psi_0^{\,2})\, D_p\, t/W^2 \right\}} \right] \qquad \cdots \quad (4.14.10)$$

But Q_{Re} the reclaimable minority charge coming out of the emitter contact is

$$Q_{Re} = \int_0^{\infty} -\,i_{pe}\, dt \qquad \cdots \quad (4.14.11)$$

$$= \frac{2I_{pe}}{(\theta + \psi_0 \coth \psi_0)} \cdot \frac{W^2}{D_p} \sum_{1}^{\infty} \frac{n^2\pi^2}{(n^2\pi^2 + \psi_0^{\,2})^2} \qquad \cdots \quad (4.14.12)$$

Summing the series using standard results given in foot-note, equation (4.14.12) gives

$$Q_{R_e} = \frac{I_{p_e} W^2}{2D_p} \left[\frac{\left(\frac{\coth \varphi_o}{\varphi_o} - \operatorname{cosech}^2 \varphi_o \right)}{\theta + \varphi_o \coth \varphi_o} \right] \quad \cdots \quad (4.14.13)$$

For the minority charge that is coming out of the collecting ohmic contact we proceed along the same lines to find out

$$i_{p_e}(t) = -AD_p q \left(\frac{dp}{dx}\right)_{x=W}$$

$$= -2I_{pe} \frac{e^{\theta}}{(\theta + \varphi_o \coth \varphi_o)} \sum_1^{\infty} \left[\frac{n^2 \pi^2}{(n^2\pi^2 + \varphi_o^2)} \cdot e^{\left\{ -(n^2\pi^2 + \psi_o^2) \frac{D_p t}{W^2} \right\}} \cdot \cos n\pi \right]$$

$$\cdots \quad (4.14.14)$$

$$\text{and } Q_C = \int_0^{\infty} i_{pe}\, dt = \frac{I_{pe} - e^{\theta}}{(\theta + \varphi_o \coth \varphi_o)} \cdot \left\{ \frac{W^2 \operatorname{cosech} \varphi_o}{2D_p \varphi_o} (\varphi_o \coth \varphi_o - 1) \right\}$$

Using equation (4.11.2)

$$Q_C = I_{re} \frac{W^2}{2D_p} a_o \left[\frac{\varphi_o \coth \varphi_o - 1}{\varphi_o^2} \right] \quad \cdots \quad (4.14.15)$$

At this stage we are in a position to deal with the origin of the capacitances in the admittance expression (4.13.9).

From equation (4.14.13) we can write, with eqn. (4.13.10)

$$\frac{\delta Q_{R_e}}{\delta V_e} = \frac{1}{r_e} \frac{W^2}{2D_p} \left[\frac{\left(\frac{\coth \varphi_o}{\varphi_o} - \operatorname{cosech}^2 \varphi_o \right)}{(\theta + \varphi_o \coth \varphi_o)} \right] \quad \cdots \quad (4.14.16)$$

Comparing (4.14.16) with (4.13.10) we conclude that it is the variation of the emitter contact reclaimable charge with voltage that is the origin of the capacitance in the input admittance expression (4.13.9).

The useful standard results are : $\Sigma \frac{1}{n^2 + a^2} = \frac{\pi}{2a} \coth \pi a$ and $\Sigma (-1)^n . \frac{1}{n^2 + a^2} = \frac{\pi}{2a} \operatorname{cosech} \pi a$ that may be reduced by differentiation with respect to a, etc. to sum the series in (4.14.12) and (4 14.14).

But this reclaimable charge is a part of the stored charge Q_s hence we find :

$$\frac{Q_{Re}}{Q_s}=\frac{W^2}{2L_p^2\,(1-\alpha_0)\,\psi_0^2}\left[\frac{\psi_0\coth\psi_0-\psi_0^2\operatorname{cosech}^2\psi_0}{\theta+\psi_0\coth\psi_0}\right]\quad\cdots\quad(4.14.17)$$

as the part of Q_s that is manifesting as due to an input capacitor C_{de} while the stored charge is due to finite life-time of minority carriers because infinitesimal life-time implies immediate recombination and no charge storage. Capacitor arising from charge storage is known as diffusion capacitor.

Similarly, we can find out the transfer admittance applicable to the study of transistor and defined by

$$y^*_{21}=\frac{i_c}{v_e}=\frac{-i_{pc}}{v_e}\quad\cdots\quad(4.14.18)$$

with the sign convention that positive current is one that enters the ohmic contact from terminal that is the collector current $i_c=-i_{pc}$.

From equation (4.10.7) with (4.10.3), we have for

$$y^*_{21}=-\frac{qI_{pe}}{kT}\cdot e^{\theta}\left[\frac{\psi\operatorname{cosech}\psi}{(\theta+\psi_0\coth\psi_0)}\right]\quad\cdots\quad(4.14.19)$$

which can be written, on Taylor series expansion around $s=0$ *i.e.*, $\psi=\psi_0$ with equation (4.13.1) and (4.11.2) as

$$y^*_{21}=\frac{1}{r_e}\left[-\alpha_0+s.\ \frac{W^2\alpha_0}{2D_p}\left(\frac{\psi_0\coth\psi_0-1}{\psi_0^2}\right)\right]\quad\cdots\quad(4.14.20)$$

Since $\phi_0\cosh\psi_0 \geqslant \sinh\psi_0$ always, so $\phi_0\coth\psi_0\geqslant 1$ hence the susceptance in y^*_{21} is capacitive and this capacitance is

$$C_{dc}=\frac{W^2\alpha_0}{2D_p r_e}\left[\frac{\psi_0\coth\psi_0-1}{\psi_0^2}\right]\quad\cdots\quad(4.14.21)$$

which can also be written using equation (4.13.1) in equation (4.14.15) as

$$C_{dc}=\frac{\delta Q_e}{\delta V_e},\quad\cdots\quad(4.14.22)$$

pointing to the fact that the capacitance in the transfer admittance is also arising from the variation of the reclaimable charge from the collector contact with emitter voltage. The fraction of the stored charge that is reclaimable from the collector contact when the emitter voltage varies is from equations (4.14.15) and (4.14.6)

$$\frac{Q_{Rc}}{Q_s}=\frac{W^2 a_0}{2L_p^2\,\psi_0^2\,(1-a_0)}(\psi_0 \coth\psi_0-1) \quad \cdots \quad (4.14.23)$$

For example, for the diffusion limited $(\theta=0)$ narrow base $(W \leqslant L_p)$ diode the transit time

$$T_B=T_p(1-a_0)=W^2/2D_p \quad \cdots \quad (4.14.24)$$

while total diffusion capacitor arising from stored charge is

$$C_{d(\text{total})}=\frac{dQ_s}{dV_e}=\frac{T_B}{r_e}=\frac{W^2}{2D_p r_e} \quad \cdots \quad (4.14.25)$$

$(\because \quad Q_s=I_{pe}T_B)$

For this particular case, we have from (4.13.12)

$$C_{de}=W^2/3D_p r_e \quad \ldots \quad (4.13.12)$$

while expanding coth ψ_0 and cosech ψ_0 in Taylor series around $\psi_0=0$ and retaining only two terms of equation (4.14.17) we have

$$\frac{Q_{Rc}}{Q_s}\simeq\frac{W^2}{2L_p^2(1-a_0)}\times\frac{\frac{1}{\psi_0}\left(\frac{1}{\theta_0}+\frac{\theta_0}{3}\right)-\left(\frac{1}{\theta_0}-\frac{\theta_0}{5}\right)^2}{\theta_0\left(\frac{1}{\theta_0}+\frac{\theta_0}{3}\right)}=\frac{2}{3} \quad \cdots \quad (4.14.26)$$

since $(1-a_0)\simeq W^2/2L_p^2$.

And from equation (4.14.21) with $a_0=1$

$$C_{dc}\simeq\frac{W^2}{2D_p r_e}\frac{\left[\psi_0\left(\frac{1}{\psi_0}+\frac{\psi_0}{3}\right)-1\right]}{\psi_0^2}=\frac{W^2}{6D_p r_e} \quad \cdots \quad (4.14.27)$$

while from equation (4.14.23)

$$\frac{Q_{Rc}}{Q_s}=\frac{1}{3} \quad \cdots \quad (4.14.28)$$

Equation (4.14.26) and (4.14.28) show that in the diffusion limited narrow base diode 2/3 of the stored charge is reclaimable from the emitter contact and the rest 1/3 is reclaimable from the collector contact. This is why of the total diffusion capacitance in equation (4.14.25) of $W^2/2D_p r_e$, a (2/3) appears as C_{de} in equation (4.13.12) as $W/3D_p r_e$ and the rest 1/3 appears as

$C_{dc}=W^2/6D_p r_e$ (of equation 4.14.27).

From equations (4.14.25), (4.13.12) and (4.14.27) we can write

$$r_e C_{dc} = T_B - r_e C_e \qquad \cdots \quad (4.14.29)$$

This is true for the diffusion limited narrow base diode. For the diffusion limited wide base diode ($\theta = 0$ and $W/L_p \geqslant 1$) we have with $\coth \frac{W}{L_p} = 1$ and $\operatorname{cosech} \frac{W}{L_p} \to 0$ from equations (4.14.17) and (4.14.23) since $a_0 \to 0$

$$\frac{Q_{Re}}{Q_s} = 1/2 \text{ and } \frac{Q_{Rc}}{Q_s} \to 0 \qquad \cdots \quad (4.14.30)$$

while from equation (4.14.7), (4.13.3) and (4.14.21) since

$$T_B \to T_p, \ C_{de} = \frac{T_p}{2r_e} \text{ and } C_{dc} \to 0 \qquad \cdots \quad (4.14.31)$$

showing that in this case half of the stored charge is reclaimable from the emitter contact and is responsible for the factor

$1/2$ in $C_{de} = T_p/2r_e$, since $\frac{dQ_s}{dV_e} = \frac{T_B}{r_e} = \frac{T_p}{r_e}$

while none is reclaimable from the collector contact, meaning thereby that $C_{dc} = 0$ and that the rest of the stored charge has gone in recombination.

Equation (4.14.17) and (4.14.23) have been represented graphically in Fig. 4.14.1 for different values of W/L_p with θ as a parameter. From this figure it is to be noted that

(1) higher the built-in field less is Q_{Re}/Q_s, hence C_{de} and higher is the frequency for which the diode is usable.

(2) Lesser W/L_p is, lower is C_{de} hence higher is the diodes usable frequency limit even for values of $\theta > 1$ (c.f. equation 4.13.8).

(a) It would be noted from Fig. 4.14.1 that for any value of θ, $(1-b)$ increases with W/L_p.

(b) And for any value of W/L_p, $(1-b)$ increases with θ.

The reason for this is obvious and that a greater length of diode causes greater recombination during switching transient.

For the last point, charges near the emitter end, to start with, were travelling towards the collection end but later in the transient period thay start travelling towards the emitter thus taking longer time for their expulsion causing more recombination.

(3) For values of $W/L_p \leqslant 1$, most of the stored charge is reclaimable through the two contacts and the built-in field makes very little difference from the no built-in field case. This is the case with practical diodes.

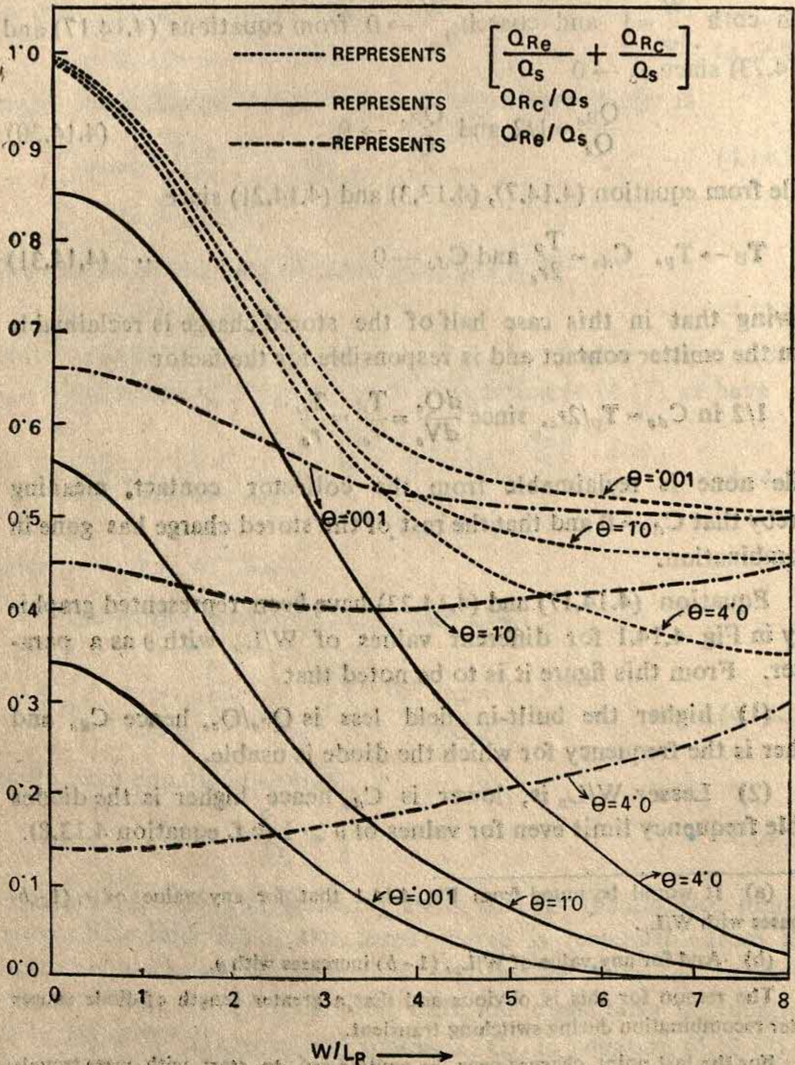


Fig. 4.14.1. Fractional reclaimable charges as related to the base field and base width

(4) For values of $W/L_p \gtrsim 1$, all of the stored charge is not relaimable through the two contacts and in general we can write

$$b = \frac{Q_{Re}}{Q_s} + \frac{Q_{Rc}}{Q_s} \quad \cdots \quad (4.14.32)$$

where $(1-b)$ is the fraction of the stored charge which is lost in recombination during the switching transient.

Because of the first three reasons the practical diodes are made narrow base only, for high frequency as well as high speed purposes.

From equation (4.14.32)

$$b.\frac{dQ_s}{dV_e} = b.\frac{T_B}{r_e} = \frac{dQ_{Re}}{dV_e} + \frac{dQ_{RC}}{dV_e} \quad \cdots \quad (4.14.33)$$

or $$r_e C_{dC} = bT_B - r_e C_{de} \quad \cdots \quad (4.14.34)$$

while for a practical diode $b \simeq 1$, hence

$$T_{21} = T_B - T_{11} \quad \cdots \quad (4.14.35)$$

where $$r_e C_{dC} = T_{21} \text{ and } r_e C_{de} = T_{11} \quad \cdots \quad (4.14.36)$$

Lastly, the two approximate forms of intrinsic admittances that we have obtained can be written from (4.13.9) and (4.14.20) with (4.14.36) as

$$\left.\begin{aligned} y^*_{11} &= \frac{1}{r_e}(1 + j\omega T_{11}) \\ y^*_{21} &= \frac{1}{r_e}(-a_0 + j\omega T_{21}) \end{aligned}\right\} \quad \cdots \quad (4.14.37)$$

4.15 Switching Characteristics of P-N Junction diode

We are now in a position to think about the transient behaviour that may be expected of a P-N junction diode. In the last article we have seen that when forward current flows through the diode, a stored charge equal to the current transit time product is stored in its body. Obviously, if the diode is switched to reverse bias as shown in Fig. (4.15.1*a*), the stored charge has to be expelled out completely before the current stops. As shown in Fig. (4.15.1b), when the diode was forward biased, the current that was flowing through the diode was $V_f/R_f = I_f$. Say at time $t=0$, the diode has been reverse biased.

During forward bias there was a certain level of minority carriers at transition region boundary, which was maintained by a voltage $V_a = \frac{KT}{q} \log \frac{p_{no}}{\bar{p}_{no}}$, that dropped across the junction. In article (4.14) we assumed that as soon as the forward voltage was withdrawn the excess minority carrier density dropped to zero.

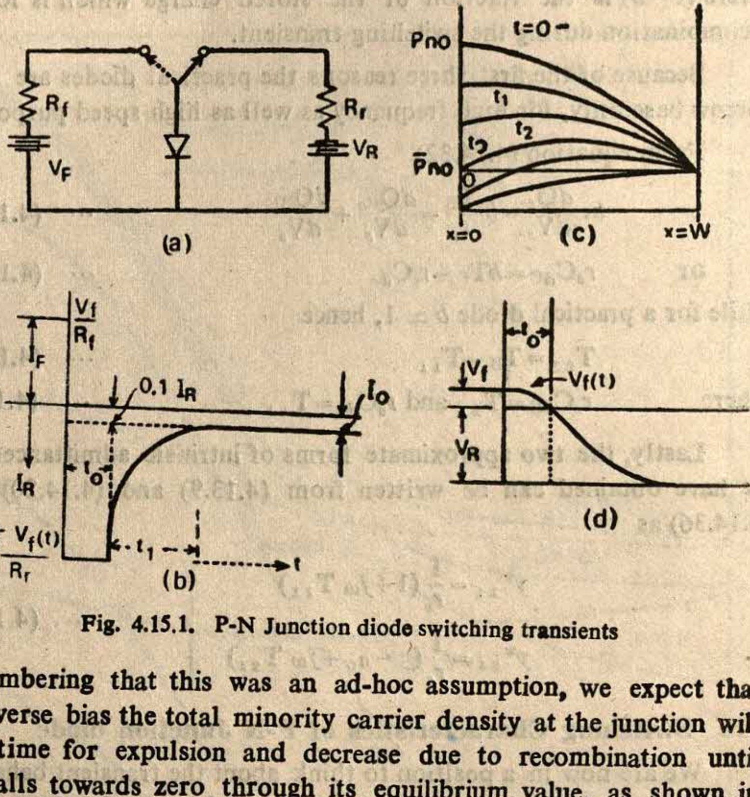


Fig. 4.15.1. P-N Junction diode switching transients

Remembering that this was an ad-hoc assumption, we expect that on reverse bias the total minority carrier density at the junction will take time for expulsion and decrease due to recombination until this falls towards zero through its equilibrium value, as shown in Fig. (4.15.1c). So long the minority carrier density at the junction boundary is above the thermal equilibrium value, the junction will behave as if it is forward biased and the voltage drop across it will be $V_f(t) = \frac{KT}{q} \log \frac{p_{no}(t)}{p_{no}}$. This period is called the charge storage phase. During this time the reverse current will be given by $\left(\frac{V_R - V_f(t)}{R_r}\right)$ which is nearly constant, (since $V_f(t)$ is of the order of a fraction of a volt) as shown in Figs. (4.15.1b and d). During this time continuous charge expulsion was occurring through both ends of the diode and this continues even after $V_f(t)$ has become zero.

Equations (4.14.10) and (4.14.14) indicate that the current expulsion occurs due to multiple time constants of which the largest time constant will persist longest. Hence we expect the reverse current of the diode to show an exponential tail, and thus the current decay is nearly exponential as shown in Fig. (4.15.1b). Even after the junction boundary excess charge has become zero the expulsion of charge continues until the boundary concentration is near zero and the full reverse voltage is established across the diode and only leakage current is flowing through it. In this we should remember that while reverse voltage is getting established the junction boundary shifts which speed up the expulsion of charges near the junction boundary causing the slopes of the junction voltage and current transients after t_0 (see Fig. 14.15.1b & d) faster.

The switching time of the diode is defined by (t_0+t_1) where t_0 is the steady current (I_R) phase time and t_1 is the fall time to 10% of this value. In a diffusion limited narrow base diode the charge stored is $\frac{W^2}{2D_p} I_F$ (c.f. equation 4.14.24) and the withdrawal is faster, larger I_R is, hence an optimistic fall time is

$$(t_0+t_1) \simeq \left(\frac{W^2}{2D_p}\cdot\frac{I_F}{I_R}\right) \qquad \cdots \quad (4.15.1)$$

This time can be even lower than this in drift diode (c.f. example at the end of art. 4.13) and for $\theta = 4$ it can be given as

$$(t_0+t_1) \simeq \frac{W^2}{64D_p}\left(\frac{I_F}{I_R}\right) \qquad \cdots \quad (4.15.2)$$

4.16 Charge Storage and Fast Recovery Diodes

For both of these high speed diodes W (the base width) is made as small and the doping profile as good as the technology can achieve in conformity with reverse voltage requirement being obtained. The material Si in the charge storage diode is so chosen that the minority carrier life time is long (of the order of a microsecond and longer) such that in the storage phase of the transient switching period almost all of the stored charge is expelled out. This is achieved by doping the slice from both ends thus having a built-in field large on the ends on either side of the transition region *i.e.*, virtually slightly slowing down the forward injection process. The net result is that the near exponential decay part of the time is very short and is of the order of pico seconds resulting in fast falling pulse edge which is rich in harmonics. This diode has got other names, such as step recovery diode or snap-back diode. These diodes, therefore, find their use amply in harmonic generators and pulse formers used with transmission lines and cavities.

In fast recovery diode, on the other hand, the storage phase is made very short by introducing recombination centers in the diode body. In order that the recombination centers be effective in the charge trapping, the diode body is at least of length L_p (hence charge storage is of the order of $I_F T_p$) and switching recovery time is proportional to the life time of minority carriers. Of course, this method of introduction of recombination centers to decrease the carrier life time cannot be carried on indefinitely since the reverse current—which is generation current, increases—causing deterioration of the reverse blocking property of the diode. For fast recovery diode materials like GaAs, which are direct band gap semiconductors, are better suited. In direct band gap semiconductors in general, the minority carrier lifetimes are much shorter and so leads to ultra high speed switching. In Si practical recovery times of the order of 1 to 5 n sec. and in GaAs of the order of 0.1 n second are achieved.

Another variation of the fast recovery diode is a metal *n*-type semiconductor diode of the Schottky type. This is a majority carrier diode with very little minority charge storage. Secondly, the carrrier densities in metal being very large the relaxation time is very small. In this connection S.C.L. diodes are also to be mentioned whose speed is considerably faster than the P-N junctions. We shall discuss these after we have studied the metal semiconductor contacts.

4.17 Avalanche breakdown and Zener breakdown

The P-N junction reverse bias breakdown is basically of two types, (1) the zener breakdown and (2) the avalanche breakdown. We here discuss the avalanche breakdown first. Under reverse bias condition the field in the transition region increases and at large field strengths ~ 200 KV/cm, collision ionisation sets in. When collision ionisation sets in, one or a few carriers may have started the process but initial ionisation created hole electron pairs are also accelerated and gain enough energy to join in the process of further ionisation.

Thus this continnes and, in a very short time, the number of carriers created become very large and hence the word avalanche is used.

The growth of the number of carriers produced can be visualised by the following argument. Say, one carrier produced m number of carriers by collision. Then $(1+m)$ number of carriers produced $(1+m)m$ number of carriers in the second round and so on, thus giving in a short time

$$1+m+m^2+m^3+\cdots\cdots=\frac{1}{1-m} \qquad \cdots \quad (4.17.1)$$

number of carriers. If m is at least one then the total number of carriers are very large indeed. Of course, if m is greater than unity then the series is a divergent one and could not be summed as above. But that does not deter us from visualising the phenomena now.

For one sided abrupt junction the maximum field occurs at the junction (c.f. Fig. 3.1.1d). Breakdown will obviously start at the maximum field region. The breakdown field E from eqn. (3.1.3) with $x=0$, and equation (3.1.6)

$$E_b=\left[\frac{2q\,N_d\,V_t}{\varepsilon\varepsilon_0}\right]^{1/2} \qquad \cdots \quad (4.17.2)$$

hence the breakdown voltage for one sided abrupt junction

$$V_b=\frac{\varepsilon\varepsilon_0\,E_b^{2}}{2q\,N_d} \qquad \cdots \quad (4.17.3)$$

Similarly, the breakdown voltage for linearly graded junction from equation (3.1.15) is

$$V_b=\frac{4}{3}\left[\frac{2\varepsilon\varepsilon_0}{bq}\right]^{1/2}\cdot E_b^{3/2} \qquad \cdots \quad (4.17.4)$$

The approximation that we are using is that for a given semiconductor the breakdown field E_b has a fixed value. This breakdown field E_b is expected to depend on the band gap, increasing with the increase of band gap. This is because breakdown starts when a large number of carriers are brought from the valence band to conduction band. A detailed calculation gives for one sided abrupt junction

$$V_b=60\left(\frac{E_g}{1.1}\right)^{3/2}(10^{16}/N_d)^{3/4}\text{ volts} \qquad \cdots \quad (4.17.5)$$

where E_g is in electron volts.

The breakdown field also depends on the radius of curvature of corners of junction. For a plane junction in silicon it is about 2.5×10^5 volts/cm and in Ge 1.18×10^5 V/cm at room temperature. Also the breakdown voltage increases with temperature. This is reasonable because the hot carriers gain energy from the field while the mean free path *l*, shortens (c.f. art. 2.11) hence more energy is lost to cross the transition region of the junction. It turns out that the avalanche diodes have positive temperature coefficient of 4*m*V/°C for 8.2 volt range diodes which is exactly double of (−2 *m*V/°C) for the cut-in voltage of a forward bias diode. Hence one avalanche

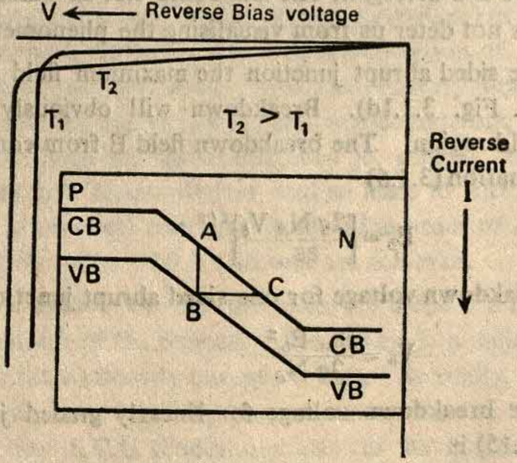


Fig. 4.17.1. Reverse breakdown I/V characteristics. The inset shows tunnelling of electrons from B to C

diode with two forward biased diodes in series gives a very accurate temperature independent reference voltage with temperature coefficient less than 0.002% per degree celsius. The diode breakdown characteristics is shown in Fig. (4.17.1). The breakdown region current can be approximated by an empirical relation of the form

$$I = \frac{I_0}{\left[1 - \left(\frac{V_a}{V_b}\right)^m\right]} \qquad \cdots \quad (4.17.6)$$

where I_0 is the reverse leakage current, V_a the applied voltage V_b, the breakdown voltage and *m* is a constant equal to 3 for Silicon and Germanium.

Zener breakdown on the other hand occurs when with reverse bias, the electrons in the depletion region of the valence band of the *p* region gain sufficient energy to directly make a transition in the conduction band of *n* region, known as tunneling (see inset of Fig. 4.17.1). To make such a transition, electrons have to either mount a potential barrier energy height AB or penetrate a potential barrier width BC. If the electrons penetrate through the barrier, we speak of tunneling of electrons to occur. The smaller this thickness is, the higher is the probability of tunneling. On the other hand, lower the transition region width smaller would be this thickness. Small transition region width occurs with higher doping. Hence there is a limitation on doping level above which tunneling will occur. This limitation makes tunnel caused start of breakdown to occur for diodes having breakdown voltage less than $(6\,E_g/q) =$ 5 volts in Ge diode and 7.5 volts in Si diode. In Si or in Ge, around these values there can be some uncertainty and normally some margin is accepted. For example, 4.8 volt zerners have tunneling started breakdown while 8.2 volt are avalanche diodes. It is to be remembered that though tunnelling starts the breakdown mechanism, ultimately avalanche takes over and avalanche occurs in all cases of breakdown of P-N junction diodes.

Zener hence tunnel caused breakdown voltage has negative temperature coefficient. This is expected since increased temperature causes the particles to have higher energies, hence at lesser reverse bias, they obtain the required probability of crossing the barrier width.

4.18 P-I-N Diode

A P-I-N diode is a P-N junction with a high resistive, either intrinsic (i) or slightly *n*-type (ν) or *p*-type (π) layer sandwiched between *p* and *n* type materials. The impurity distribution, space charge and field distribution in *p-i-n* and *p-π-n* diodes are shown in Fig. (4.18.1). Because of very low doping the entire *i*-region is usually depleted even at zero bias. Under reverse bias the depletion region extension does not increase appreciably because of heavy *p* and *n* dopings on either side and most of the voltage drops in

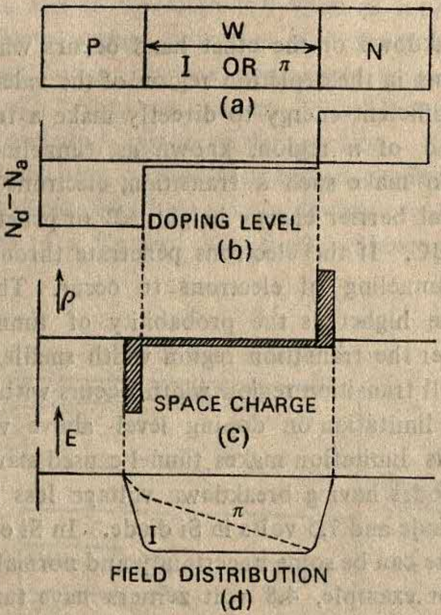


Fig 4.18.1. (a) P-I-N Diode, (b) Doping level, (c) Space charge and (d) Filled variation in it are shown.

the *i*-region with consequent increase of field there. Therefore, the reverse bias capacitance decreases little from its unbiased value as shown in Fig. (4.18.2). Since the variation is small so, (C/C_o) appears linear over the voltage range.

As the reverse bias increases the series resistance decreases rapidly towards an asymptotic value corresponding to the contact resistance plus the resistance of the substrate as shown in Fig. (4.18.2). The breakdown voltage V_{br} of a *p-i-n* diode is approximately $(W.E_m)$ where E_m is

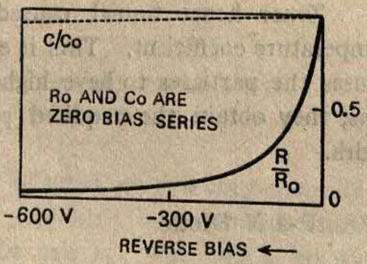


Fig 4.18.2. Shows variation of reverse bias capacitance and resistance normalised to zero bias values

the maximum electric field at breakdown. For Si, $E_m \sim 2 \times 10^5$ V/cm hence for

$$W = 30 \text{ microns} \quad V_{br} = 600 \text{ volts.}$$

Under forward biased condition holes and electrons are injected by the *p* and *n* regions into the *i*-region. The current crossing from *p* to *i* is entirely hole current, similarly that crossing from *n* to *i* is entirely electron current. The *i* region is flooded with electrons and holes so the conductivity of the *i*-layer is greatly increased in proportion to the injection level hence proportional to the d.c. current flowing (c.f. equation 4.12.6). Hence the resistance of the *i*-layer will be inversely proportional to forward current (I_F).

Because of these forward and reverse characteristics the *p-i-n* diode finds wide application in Microwave circuits. It can be used as a MW switch with essentially constant depletion layer capacitance and high power controlling capability. The switching speed is approximately ($W/2v_s$) where v_s is the saturation velocity in the intrinsic region. In addition a *p-i-n* diode can be used as a variable attenuator (variolosser) by controlling device resistance under forward bias by varying forward current. It can also be used as a modulator to perform signal modulating functions upto GHz range.

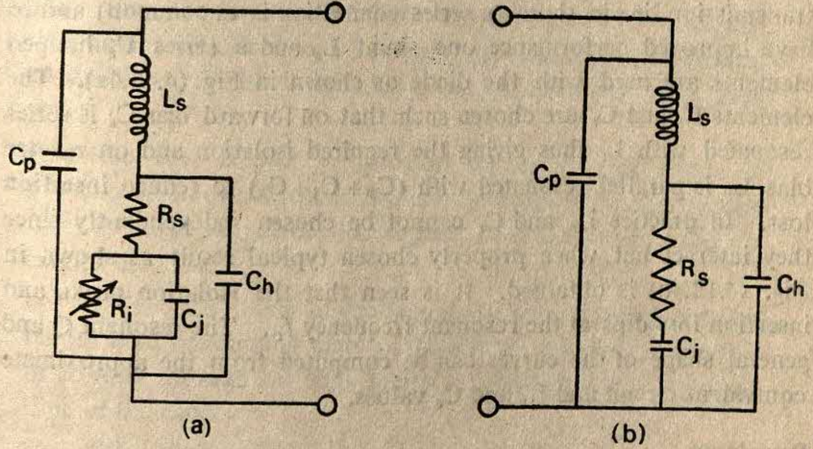


Fig. 4.18.3. Equivalent ciruit of P.I.N. diode. (a) Forward bias, (b) Reverse bias case

The approximate equivalent circuits under forward and reverse bias are shown in Fig. (4.18.3). In this figure the subscripts s for series, *j* for junction, *h* for header and *p* for parasitic are used.

To use the diode as microwave switch it is connected across the transmission line in shunt (a series connection is uncommon) and to have improved performance one shunt L_p and a series C_s lumped elements are used with the diode as shown in Fig. (4.18.4a). The elements L_p and C_s are chosen such that on forward bias C_s is series resonated with L_s thus giving the required isolation and on reverse bias L_p is parallel resonated with $(C_p + C_j + C_h)$ to reduce insertion loss. In practice L_p and C_s cannot be chosen independently since they interact but when properly chosen typical result as shown in Fig. (4.18.4b) is obtained. It is seen that the isolation peaks and insertion loss dips at the resonant frequency f_0. The resonant Q and general shape of the curves can be computed from the approximate equivalent circuit and L_p and C_s values.

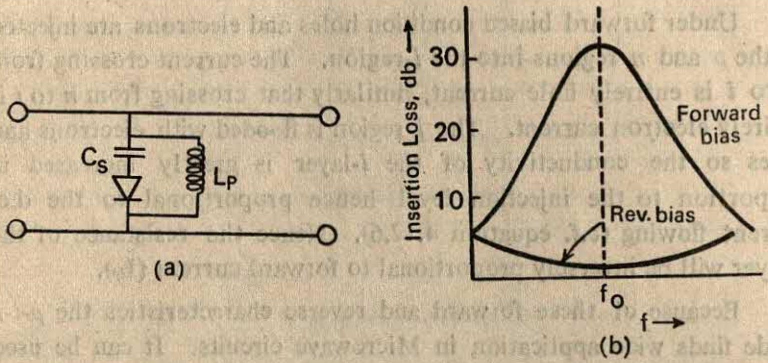


Fig. 4.18.4. (a) Transmission line with diode, (b) Typical performance of shunt P.I.N. switch

Problems

4.1. Calculate the reverse saturation current of a P-N junction narrow base diode doped to level 10^{17}/c.c. in the emitter bulk of N-type and 10^{15}/c.c. in the base body. Assume a diode diameter of 10 microns $D_n = 50$ cm^2/sec and a life time $T_n = 0.5$ microsecond. Give reasons for finding experimental values of leakage current orders of magnitude larger than this calculated value though the diode parameters are typical ones.

[Ans. 2.826×10^{-16} Amp.]

4.2 Explain how leakage current versus reciprocal of teemperature T plot can give band gap of the material of the diode.

4.3. Calculate the transport factor a_o for a narrow base diode $W = 0.1\ L_p$ and an wide base diode $W = 2L_p$.

Ans. ·995 and 0·27

4.4 Calculate the injection efficiency of the *n-p* junction diode of problem 4.1 above having other physical parameters as $W_p = 2$ microns, $W_n = 200$ microns and $D_p = 20\ cm^2/sec$.

4.5. Quote a numerical value for the small signal voltage across a *p-n* junction diode. Hence calculate the small signal r.m.s. current through the diode for 1 *m*A d.c. current.

4.6. Calculate at 1 MHz the a.c. impedance of a *p-n* junction silicon diode at 1 *m*A having a base region built in field of 200 v/cm, width 2 microns, base diffusion length 20 microns and minority carrier life time 2 microseconds.

4.7. Find the a.c. impedance of the diode of problem 4 above if the built in field is zero. Compare the values of the diffusion capacitances. What is its value if the diode was a wide base one ?

4.8. Find the low frequency transport factor for the diode of problem 4.6 above. Hence calculate the minority carrier transit time through the diode. Compare its value with minority carrier life time.

4.9. The *p-n* junction diode of problem 4.6 above is switched from its forward conducting state of 10 *m*A to a back bias voltage of 20 volts through 1 k ohm resistor. Estimate the switching time of the diode.

Ans. 0.5×10^{-8} secs.

4.10. Calculate the breakdown voltage of a Si diode having emitter doping 10^{17}/c.c. base doping 10^{15}/c.c. Assume a breakdown field of 2.0×10^5 v/cm and a dielectric constant of 12.

Ans. 134 v.

4.11. Calculate the reverse bias breakdown voltage of a *p-n* junction diode of emitter region doping 5×10^{16}/c.c. while the base region doping is $5 \times 10^{16}\ \exp(-4.6x/W)$, where x is in microns. Assume a dielectric constant of 12 and base thickness of 5 microns and the breakdown field of prob. 4.10.

4.12. An n-type semiconductor of length L has a d.c. voltage V applied to the ohmic contacts at the ends. A unit impulse of minority concentration has been put in at $x=0$ and time $t=0$. Show that the concentration at a distance x at time t is given by

$$p(x, t)=\exp\left(-t/\tau_p\right) \exp\left[-(x-\mu_p \mathrm{E}t)^2/4\mathrm{D}_p t\right]/(4\pi \mathrm{D}_p t)^{\frac{1}{2}}$$

where $\mathrm{E}=(\mathrm{V}/\mathrm{L})$. Note the impulse has become a Gaussian distribution moving and spreading with time and decaying with a life time τ_p. Hence find the location of the peak of the pulse at time $t=t_0$ and its velocity as $\mu_p\mathrm{E}$.

4.13. A point contact probe at $x=d$ properly biased through a resistive load shows a voltage developed across the load if the pulse of problem (12) above is observed on an oscilloscope. Show that the time delay (δt) between the peak voltage and $(1/e)$ of the peak value is related to the minority carrier diffusion constant $\mathrm{D}_p=(\mu_p\mathrm{E})^2 \times (\delta t)^2/16\, t_0$ where $t_0=(d/\mu_p\mathrm{E})$ is large compared to (δt) and $\exp(-\delta t/\tau_p) \simeq 1$.

4.14. Show that the integral of the quantity $p(x, t) \exp(-t/\tau_p)$ of problem (12) above over the entire physical length of the semiconductor (assuming it to be large) is unity. Explain the implication of this.

4.15. Establish that natural logarithm of the pulse height versus the time t_0 of flight for different positions of the probe of problem (13) above gives a straight line of slope $(1/\tau_p)$.

Hint : The voltage is proportional to the number of carriers collected of the ones that survives recombination until the peak comes under the probe ; hence it is of the form $\exp(-t_0/\tau_p)$.

4.16. Formulate an experimental method based on problems (12) to (15) above for the measurement of minority carrier mobility, diffusion constant and life time. This elegant method is the celebrated Haynes Shockley experiment (cf. S. Wang, p. 290—294).

4.17. Illumination of a semiconductor surface generates excess carriers. If the rate of generation is G, the recombination velocity is v_s, the light is non-penetrating (all light is absorbed very near the surface) that is, the absorption coefficient is large $= 10^4$/cm and higher and the surface is an extended one, write down an expression for the dispersal of the excess minority carriers on (1) the illuminated surface at $z = b$, (2) on the unilluminated surface which is at $z = 0$, as shown in Fig. (4.17.2).

(1) $D_p \dfrac{dp_1}{dz} = G - v_s p_1$ at $z = b$. (2) $D_p \dfrac{dp_1}{dz} = v_s p_1$ at $z = 0$.

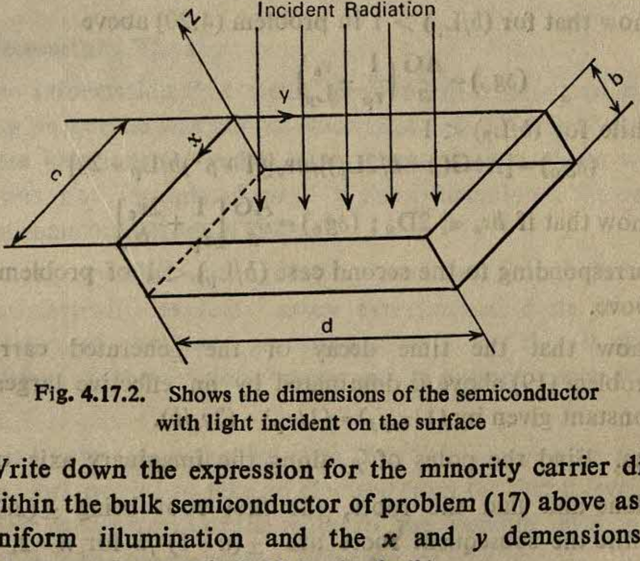


Fig. 4.17.2. Shows the dimensions of the semiconductor with light incident on the surface

4.18. Write down the expression for the minority carrier dispersal within the bulk semiconductor of problem (17) above assuming uniform illumination and the x and y demensions large. (Neglect any electric field in the bulk).

Ans. $(\delta p_1/\delta t) = D_p(\delta^2 p_1/\delta z^2) - p_1/\tau_p$

4.19. Assuming that the carrier generation is caused by light pulses

$$\bar{p} = \int_0^\infty p_1 \exp(-st)\, dt \text{ and } a^2 = (1 + s\tau_p)/D_p\tau_p$$

find $\bar{p}$ along z direction in the bulk semiconductor using the boundary conditions of problem (4.17) above.

Ans. $\bar{p} = G[D_p a \cosh az + v_s \sinh az]/s[(D_p{}^2 a^2 + v_s{}^2)$
$. \sinh ab + 2v_s D_p a \cosh ab]$

4.20. Find the steady state solution of minority carrier distribution from the result of problem (4.19) above and hence the steady state change in sample conductance by

$$(\delta g_0) = \int_0^b \int_0^c (q p_1(\mu_n + \mu_p) dz.dx/a)$$

measured between the faces perpendicular to y separated by a distance d the sample width being c. with $D_p \tau_p = L_p^2$, $\beta = (D_p/L_p v_s)$ and $cq(\mu_n + \mu_p) = A$.

Ans. $(\delta g_0) = AGL_p[\cosh(b/L_p) - 1 + \beta \sinh(b/L_p)]/$
$av_s[(1+\beta^2)\sinh(b/L_p) + 2\beta\cosh(b/L_p)]$

4.21. Show that for $(b/L_p) > 1$ in problem (4.20) above

$$(\delta g_0) = \frac{AG}{a}\left(\frac{1}{\tau_p} + \frac{v_s}{L_p}\right)$$

while for $(b/L_p) < 1$

$$(\delta g_0) = [AbG(\beta + b/2L_p)]/av_s[(1+\beta^2)b/L_p + 2\beta]$$

4.22. Show that if $bv_s \leqslant 2D_p$; $(\delta g_0) = \frac{AG}{a}\left[\frac{1}{\tau_p} + \frac{2v_s}{b}\right]$ corresponding to the second case $(b/L_p) < 1$ of problem (4.21) above.

4.23. Show that the time decay of the generated carriers of problem (19) above is dominated by an effective largest time constant given by $(1/\tau_{eff}) = (1/\tau_p) + (2v_s/b)$.

Ans. Find the poles of $\bar{p}$ along the imaginary axis of the s plane and find the roots, the first root being $\pm(2v_s/D_p b)^{\frac{1}{2}}$ while the subsequent roots are $\pm(n-1)^{\pi}/b$ for $n = 2, 3, 4, \cdots$ etc. Hence find the decay time constants from the time dependent solution for $p_1(t)$ obtained by using contour integration and residue theorem.

4.24. A photon stream of average power P at mean optical frequency f incident on a semiconductor gives a photo current I. Find the ratio of average number of excited particles (electrons or holes) to the average number of incident photons per second. What would you call this ratio ? On what factors does it depend ?

Ans. Conversion or quantum efficiency.

Chapter 5

METAL SEMICONDUCTOR JUNCTION

In this chapter we shall deal with the simple theories regarding the metal semiconductor junctions as well as some special devices, such as, the Schottky diade, Varactor diode, and space charge limited diode.

5.1 Elementary Theory

The information that metal semiconductor junctions can show rectifying properties was known since 1874. In 1904 Prof. J. C. Bose used these junctions for radio wave detection, though an workable theory was not formulated until Wilson formulated the transport theory of semiconductors in 1931.

The theory that was used till the sixties, though do not work well, satisfactorily explains many experimental facts for certain combinations of metal and semiconductors. For this it would be worth discussing the previous theory and its present more improved form.

It is known to us that by work function ($\varphi_m \cdot q$) of a metal we mean the work needed to take an electron from the Fermi level (that is where their energies are) of the metal to the vacuum level, meaning thereby that it has overcome the influence of the metal. But when when we speak of a semiconductor, the above definition is not applicable since electrons reside in the conduction band and not on the Fermi surface. Hence the energy needed to take an electron from the conduction band to the so-called vacuum level is called by a different name, 'the electron affinity' ($q \cdot \chi_s$). Though the above formal definition of work function does not apply for semiconductors, still the energy difference between the vacuum level to the Fermi level is loosely used by the work function ($\varphi_s \cdot q$) of a semiconductor. Hence for semiconductor

$$\varphi_s = \chi_s + (E_C - E_F)/q \qquad \cdots \qquad \cdots \qquad (5.1.1)$$

Depending on the relative values of the work functions of metal and semiconductor and that a semiconductor can be of two types *n* or *p*, four distinct cases can arise. These are as follows :

(a) Metal and *n*-type semiconductor having $q\varphi_m > q\varphi_s$

Fig. (5.1.1a) and Fig. (5.1.1b) show the energy level diagrams of the metal and semiconductor before contact and after contact. On contact electrons from the semiconductor enter the metal, leaving fixed positive charges behind. Thus the semiconductor lowers its

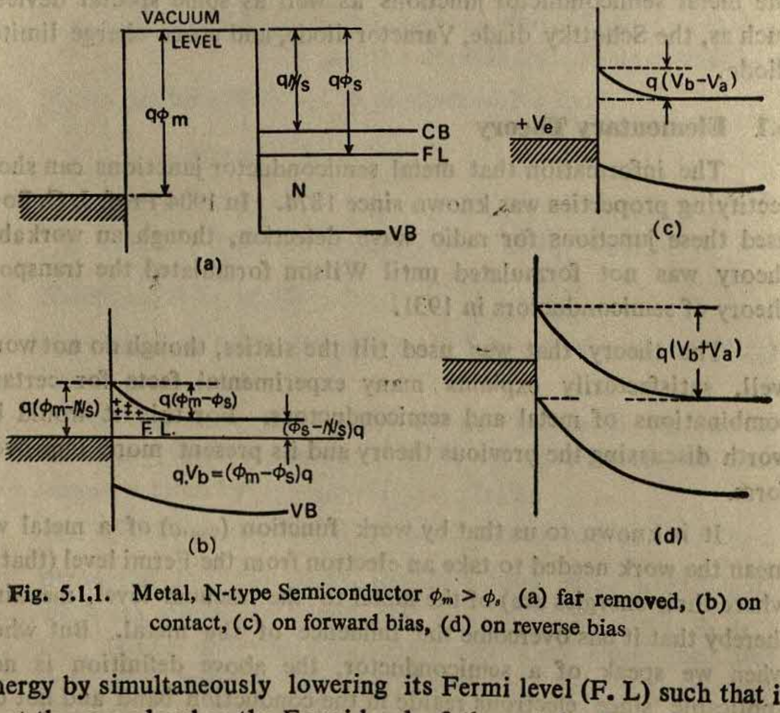


Fig. 5.1.1. Metal, N-type Semiconductor $\phi_m > \phi_s$ (a) far removed, (b) on contact, (c) on forward bias, (d) on reverse bias

energy by simultaneously lowering its Fermi level (F. L) such that it is at the same level as the Fermi level of the metal. This brings in a field which becomes as much strong as is needed to stop further charge migration and an equilibrium is reached. Deep within the body of the semiconductor the difference in energy between the Fermi level and the conduction or valence bands remain unchanged. At the contact the F. L. is lowered by $q\,(\varphi_m - \varphi_s)$ resulting in the diffusion potential $(\varphi_m - \varphi_s)$. The positive fixed distributed charge in the semiconductor holds a sheet of charge of electrons on the metal giving the total charge neutrality. The direction of the field at the contact is negative. The space charge region thickness depends on the donor concentration and the junction can be regarded as one sided abrupt, under equilibrium condition the net current crossing the junction is zero.

If the metal is made positive with respect to the semiconductor the barrier height and width are reduced as shown in Fig. (5.1.c) and if there is unlimited supply of electrons in the conduction band of the semiconductor, then the current can increase with voltage continuously. If, on the other hand, the metal is made negative with respect to the semiconductor, then the barrier height and width increases as shown in Fig. (5.1.d). Since the concentration of electrons at the contact in the conduction band is $N_c \exp[-q(\phi_m - \psi_s)/kT]$ and is fixed, therefore, for metal semiconductor combinations for which $q(\phi_m - \psi_s)$ is appreciably large (0.3. ev) the current will be limited. Therefore, this contact will be called a rectifying contact.

(b) Metal-*n*-type semiconductor $q\phi_m < q\phi_s$

Figs. (5.1.2a) to (5.1.2d) show the energy bands in metals and semiconductor before contact, after contact, metal side positive biased with respect to the semiconductor side and the metal side negative with respect to the semiconductor respectively.

Before contact the metal Fermi level was at higher energy than that of the semiconductor, so electrons spill from the metal to the semiconductor on contact. The negative space charge at the semiconductor side of the contact keeps itself bound to the positive sheet charge on the metal surface. A field is brought into existence which becomes ultimately strong enough to stop electron migration, the two Fermi levels are equalised and equilibrium is established. Thus the accumulation of mobile electron charge occurs on the semiconductor at the contact, the magnitude of which is larger than that in the bulk semiconductor and this accounts for the closeness of the CB to the FL in the semiconductor at the contact than in the bulk. Here the field is positive, *i.e.*, directed from metal to semiconductor; check that the corresponding potential diagram would also give a positive field. The built in junction potential is $(\phi_s - \phi_m)$ in this case. Fig. (5.1.2c) shows the case when metal is made positive with respect to the semiconductor. If there is an unlimited supply of electrons in the conduction band of the semiconductor they can roll down the energ gradient to the metal hence theoretically no limitation of current flow may occur. On the other hand, when the semiconductor is made positive with respect to the metal (c.f. Fig. 5.1.2d); since at the

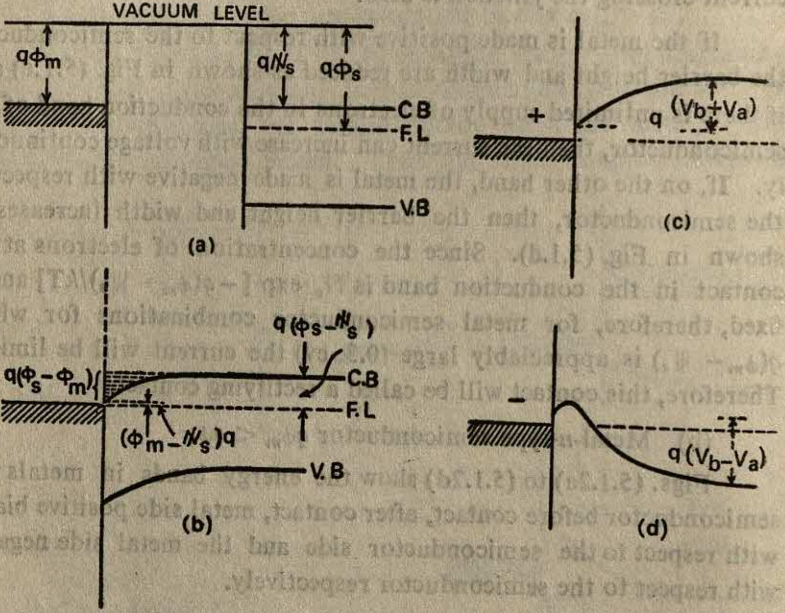


Fig. 5.1.2. Metal, N-type semiconductor $\phi_m < \phi_s$: (a) before contact, (b) on contact, (c) metal positive biased w.r.t sem, (d) metal negative biased w.r.t. sem

contact within the semiconductor the concentration of carriers is $N_o \exp{-q(\phi_m - \chi_s)/kT}$, which is large, [in the extreme case $(\phi_m - {}_s\chi)$ can be negative causing degeneracy of the semiconductor at the contact], large current can flow through the contact with increase of voltage. That is we see that such a contact does not produce any rectification nor does it interfere in current flow, hence it is called an ohmic contact, because the current for very small applied voltages will follow Ohm's law, the ohmic resistance being due to the semiconductor bulk. The presence of surface energy seats may lead to apparent invalidity of the above mentioned behaviour.

(c) Metal-*p*-type semiconductor $q\Phi_m > q\Phi_s$

In this case, before contact and after contact, the energy level diagrams are shown in Figs. (5.1.3a) and (5.1.3b). On making contact, electrons flow from semiconductor to metal until equilibrium is reached. The V.B. of the semiconductor comes nearer or even can cross the metal F.L. thus causing accumulation of holes or even the semiconductor can become degenerate *p*-type at the contact. From the semiconductor holes can easily roll down the potential gradient to the metal and get recombined with electron when the semiconductor side is made positive with respect to the metal. The energy level diagram for this situation is shown in Fig. (5.1.3c).

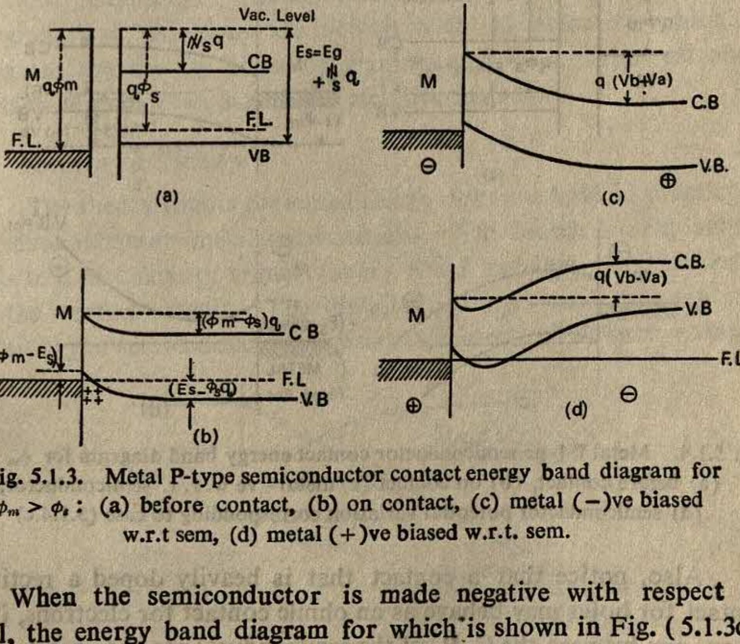


Fig. 5.1.3. Metal P-type semiconductor contact energy band diagram for $\phi_m > \phi_s$: (a) before contact, (b) on contact, (c) metal (−)ve biased w.r.t sem, (d) metal (+)ve biased w.r.t. sem.

When the semiconductor is made negative with respect to metal, the energy band diagram for which is shown in Fig. (5.1.3d), the number of holes in the V.B. of the semiconductor at the contact being large, they can easily roll down the potential gradient along the semiconductor body. Thus we see that this contact does not cause any hindrance to current flow in either bias conditions, hence it would be called an ohmic contact for holes.

(d) Metal P-type semiconductor $q\Phi_m < q\Phi_s$.

The energy level diagram in this case before and after contact are shown in Figs. (5.1.4a) to (5.1.4c), which correspond to before contact, after contact and when the semiconductor side is positive with respect to the metal. Fig. (5.1.4d) shows the potential profile corresponding to Fig. (5.1.4c) where we see that if there is enough supply of holes, they can roll down the potential gradient to the metal. For reverse bias, the student can easily visualise that there would be a limitation in the hole current since the hole concentration in the semiconductor V.B. at the contact is $N_v \exp[-(E_s - q\Phi_m)/kT]$, small since $(E_s - q\Phi_m)$ is large. Thus this contact is a rectifying contact for holes.

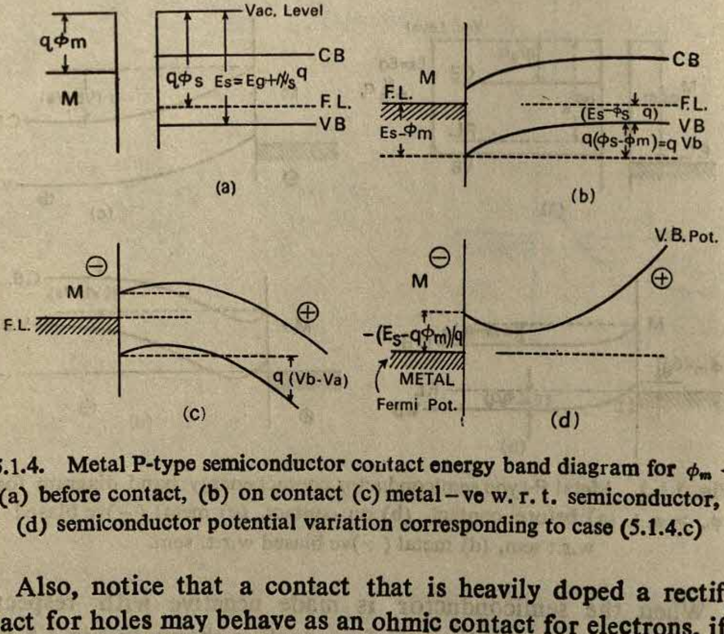


Fig. 5.1.4. Metal P-type semiconductor contact energy band diagram for $\phi_m < \phi_s$ (a) before contact, (b) on contact (c) metal – ve w. r. t. semiconductor, (d) semiconductor potential variation corresponding to case (5.1.4.c)

Also, notice that a contact that is heavily doped a rectifying contact for holes may behave as an ohmic contact for electrons, if the energy gap between the C. B. of the semiconductor to the metal Fermi level becomes small enough (cf. Fig. (5.1.4b) and (5.1.2b). Gold with P-type Ge gives ohmic contact while with *n*-type Ge gives rectifying contact.

The depletion region width of a metal semiconductor junction can easily be calculated by assuming it as an abrupt step junction if the semiconductor doping is uniform.

This would be given from equation (3.1.6) as,

$$d = \left[\frac{2\varepsilon\varepsilon_0 V_t}{q N_a}\right]^{1/2} \qquad \cdots \quad (5.1.2)$$

Similarly the capacitance from equation (3.1.8) as ;

$$C_j = \left(\frac{A\varepsilon\varepsilon_0}{d}\right) = A\left[\frac{\varepsilon\varepsilon_0}{2}\cdot\frac{qN_a}{V_t}\right]^{1/2} \qquad \cdots \quad (5.1.3)$$

It is to be noted that if N_a is very large, then the depletion region would be very narrow and for degenerate doping it can be so thin that a contact supposed to be rectifying turns out to be an ohmic contact because the depletion region width has become of the order of a few tens of angstrom and the particle tunnels through the junction. Example of this is obtained in IC fabrication where both *n*-regions and *p*-regions are coated with Al prior to which the *n*-regions are heavily doped by making it n^+ on *n*. Thus all ohmic contacts are achieved in a single metalisation step.

5.2 Improved Theory

The theory that is presented above does not hold in practice for all semiconductor metal combinations. The search for explanation leads to a satisfactory refined theory which ascribes the discrepancy on the surface states. The detailed energy band diagram of a metal-*n*-type semiconductor is discussed below. The basic points of refinement are (*i*) when metal semiconductor contact is made there remains an interfacial layer of atomic dimension which can stand potential difference but is transparent to electrons, (*ii*) there are on the semiconductor surface, seats of energy which is a property of the semiconductor concerned and are independent of the metal, (*iii*) that there is the lowering of the energy barrier height $(\Phi_m - \chi_s)\, q$ of Fig. (5.1.1b) due to Schottky effect. Schottky effect is lowering of the work function of the metal due to the field present on the metal

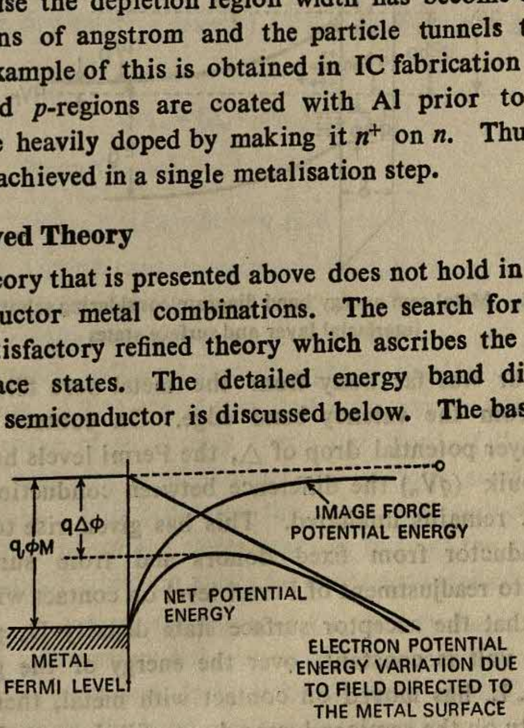


Fig. 5.2.1. Schottky contact field lowering the work function by ($\Delta\phi$)

caused by the presence of the semiconductor as shown in Fig. (5.2.1) as $(q \triangle \Phi)$. Taking these factors into account, the general metal semiconductor energy band diagram is drawn in Fig. (5.2.2) assuming that there are surface states on the semiconductor which when the semiconductor was far away from the metal were filled up to the level $q\Phi_0$ from the valency band edge. On contact there is an interfacial layer potential drop of $\triangle$, the Fermi levels have equalised but in the bulk (qV_n) the difference between conduction band edge E_c and F. L. remains unaltered. This has given rise to charges on the semiconductor from fixed donors and from surface charges exposed due to readjustment of Fermi level on contact with metal. If we assume that the acceptor surface state density D_s per unit area per electron volt is constant, over the energy of the gap, through which the F. L. has moved on contact with metal, then the surface charge density on the semiconductor due to filled up surface states is

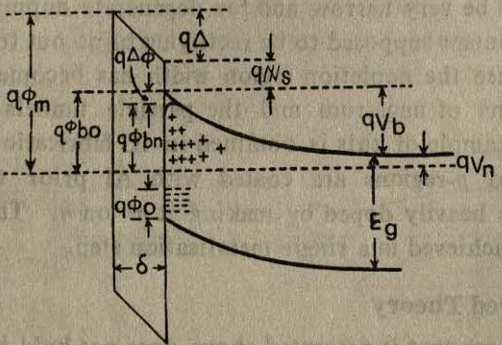


Fig. 5.2.2. Metal sem energy band diagram considering schottky effect, interfacial layer and surface states

$$Q_{ss} = -qD_s\,[E_g - q\Phi_{bn} - q\Phi_0 - q\triangle\Phi] \qquad \cdots \quad (5.2.1)$$

The depletion layer space charge assuming abrupt step junction is $Q_{so} = qN_d.\,d$, with equation (3.1.6) is given by

$$Q_{so} = [2\varepsilon_0\varepsilon_s N_d\,(\Phi_{bn}q + \triangle\Phi q - V_n q - kT)]^{1/2} \qquad \cdots \quad (5.2.2)$$

since free particle energy per dagree of freedom is $(kT/2)$.

The charge on the metal surface Q_m is, therefore

$$Q_m = -(Q_{ss} + Q_{sc}) \qquad \cdots \quad (5.2.3)$$

The potential energy drop across the interfacial layer width δ is

$$\Delta = -\delta \frac{Q_M}{\varepsilon_i \varepsilon_o} \quad \cdots \quad (5.2.4)$$

Lastly from the F.L. equality in Fig. (5.2.2)

$$\Delta = \Phi_m - (\chi + \Delta\Phi + \Phi_{bn}) \quad \cdots \quad (5.2.5)$$

Convenient and masterly manipulation of these equations gives* two results for two extreme cases of surface states (c.f. ref. 1)

(*i*) When $D_s \rightarrow \infty$

$$q\Phi_{bn} = (E_g - q\Phi_0) - q\Delta\Phi \quad \cdots \quad (5.2.6)$$

and (*ii*) When $D_s \rightarrow 0$

$$q\Phi_{bn} = q(\Phi_m - \chi) - q\,\Delta\,\Phi \quad \cdots \quad (5.2.7)$$

Note that the second case corresponds to the theory discussed in article 5.1 except for Schottky's correction of work function. The experimental results over the years of various metal and semiconductor systems indicated that straight line relationships of the form of equation (5.2.6) are obeyed and it is seen that for Si, GaAs and GeP the value of Φ_0 is very nearly one third of the band gap. Also it was seen for gold contact on various semiconductors, $(E_g - q\Phi_0) = 2E_g/3$. This indicated that most semiconductors have a high peak density of surface states near one third of the gap from the V.B. edge. Similar situation exists for other semiconductors.

For CdS, of course, the value of $q\Phi_0$ is very large (~ 0.8 E_g) and the metal CdS system behaves as if there is no surface states. Here the states are assumed to be near the band edge (CB) and the F.L. can move within the band gap over large range without affecting surface state occupancy.

---

* For well cleaned and vacuum cleaved surfaces of less than 10^{18}/cc doped semiconductors in intermediate cases, we have (cf. ref. 1) approximetely,

$$\phi_{bo} = \phi_{bn} + \Delta\phi = c(\phi_m - \chi) + (1-c)(E_g - \phi_0 q)/q$$

where $(1/c) = (1 \times q^2 \delta\, D_s/\epsilon_i)$. Using $\delta \sim 4-5°A$ and $\epsilon_i = \epsilon_o$ gives $D_s = 1.1 \times 10^{13} \times (1-c)/c$ states/cm^2/ev. Forward bias characteristics extrapolated to zero voltage gives ϕ_{bn} (cf. eqn. 5.3.13), likewise extrapolated reverse characteristic gives ϕ_{bo}. From the slope of ϕ_{bn} versus ϕ_m the constant c can be determined hence D_s. Symbols like ϕ, χ_s, ϕ_{bn} are used to designate the corresponding potentials.

5.3 Current Voltage Relation for Schottky Diode

The Schottky barrier diode is made from non-degenerate semiconductor metal combinations. On one side there having an ohmic contact and on the other the Scottky barrier. The current transport through the Schottky barrier diode is mainly due to majority carrier flow. The electron energy band diagram of the Schottky barrier diode is shown in Fig. (5.3.1). Note that electronic charge is multiplied to the corresponding potentials to state the energies. At present the current flow is known to be governed by a combination of the factors (1) the thermionic emission of electrons over the Schottky barrier, (2) diffusion of carriers in the depletion region that is particularly important near the contact region where (3) the field plays an important role. On the contact side of the potential minimum the field pulls the electrons to the metal while on the other side the field pushes the electrons away from the barrier top. Therefore near the metal contact field induced electron tunnelling may be expected. Moreover the thermionic emission depends on the velocity distribution of carriers in the barrier region while diffusion depends on the concentration gradient that may be significant in the barrier region. In addition some of the electrons, that cross the potential energy maximum (barrier) in a Schottky diode, may get back scattered due to optical phonon scattering, thus leaving a smaller net current to flow. Further the classical velocity

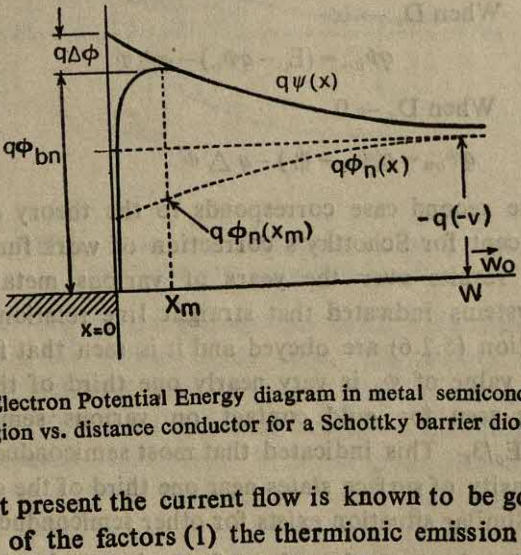


Fig. 5.3.1. Electron Potential Energy diagram in metal semiconductor barrier region vs. distance conductor for a Schottky barrier diode.

distribution of carriers get modified due to quantum mechanical reflection of electrons by the Schottky barrier and that there can be tunnelling of electrons through the barrier which again is barrier region field dependent.* Considering these we study the Schottky barrier diode from simpler considerations to impress upon the reader the essential features of some of the processes involved in the current flow in Schottky barrier diode.

5.3.1 Thermionic Emission Model

In this model the barrier height Φ_{bn} is assumed to be large compared to the particle's thermal energy which is of the order of kT, the particles having Maxwellian velocity distribution. The electron is assumed not to suffer any collision in the transit over the barrier, neither it is affected by its image in the metal moving in the opposite direction.

The electron current density from the semiconductor to metal

---

* It turns out that the fraction of particles passing a thin rectangular barrier due to tunnelling can be described by an attenuation constant α, where the wave amplitude decreases exponentially. For a barrier thickness t the transmitted fraction of amplitude is given by exp $(-\alpha t)$. This constant turns out to be $\alpha\left[\frac{2m^*}{\hbar}(V-E)\right]^{1/2}$, where V is the energy barrier height and E the total energy of the electron. In devices we frequently encounter approximate triangular barriers of finite width having a certain height V_o. In such casses the $(V-E)$ term can be written in terms of the constant field E_x of the barrier region as $(V_o - qE_x.x)$. The approximate wave amplitude decreases in passing through the barrier of width $d = (V_o/qE_x)$ can be written as

$$\left[\exp - \frac{1}{\hbar}\int_0^d \left\{2m^*(V_o - qE_x x)\right\}^{1/2} dx\right]$$

The transmission coefficient is square of this and comes to

$$\exp[-4\sqrt{2m^*}\, V_o^{3/2}/3qE_x\hbar]$$

This shows the field dependence mentioned in the text.

can be calculated using unidirectional normalised velocity distribution function* as

$$J_{sm} = qn\sqrt{m^*/2\pi kT}\int_{v_0}^{\infty} V_x \exp\left(-m^* V_x^2/2kT\right) dV_x \quad \cdots \quad (5.3.1)$$

which assumes the particles, having x directed Kinetic energies ($\frac{1}{2}m^* v_0^2$) greater than the barrier height $q(V_b - V_a)$, surmount the barrier. That is we have

$$m^* v_0^2 = 2q(V_b - V_a) \quad \cdots \quad (5.3.2)$$

and n is the concentration at the barrier that can be written as

$$n = N_c \exp\left[-(E_c - E_F)/kT\right] = N_c \exp\left(-q\Phi_{bn}/kT\right) \quad \cdots \quad (5.3.3)$$

On integration equation (5.3.1) gives

$$J_{sm} = qn\left(\frac{kT}{2\pi m^*}\right) \exp\left(-m^* v_0^2/2kT\right) \quad \cdots \quad (5.3.4)$$

which with equations (5.3.2) and (5.3.3) gives, with N_c from equation (2.10.6), for the electron flow from semiconductor to metal as

$$J_{sm} = AT^2 \exp\left(-q\Phi_{bn}/kT\right) \exp\left(qV_a/kT\right) \quad \cdots \quad (5.3.5)$$

where $A = (4\pi q m^* k^2/h^3)$ which is the Richardson constant for thermionic emission into a vacuum. For free electrons its value being 120 Amp/cm²/°K². For semiconductors m^* is the effective mass at the lowest conduction band edge.

---

* Normalised Maxmellian velocity distribution function *i.e.* the probability of an electron having its velocity lying between V_x to $V_x + \delta V_x$, V_y to $V_y + \delta V_y$ and V_z to $V_z + \delta V_z$ is given by

$$\left(\frac{m^*}{2\pi kT}\right)^{3/2} \left[\exp - m^*\left(V_x^2 + V_y^2 + V_z^2\right)/2kT\right] dV_x, dV_y. dV_z$$

This gives the fractional number of particles having velocities between V_x to $V_x + \delta V_x$ aiming for the x direction as

$$\frac{\delta n}{n} = \left(\frac{m^*}{2\pi kT}\right)^{3/2} \left[\exp\left(-m^* V_x^2/2kT\right).dV_x\right] \left[\int_{-\infty}^{\infty} \exp\left(-m^* V_y^2\right)/2kT.dV_y\right]$$

$$\left[\int_{-\infty}^{\infty} \exp\left(-m^* V_z^2/2kT\right) dV_z\right] = \left(\frac{m^*}{2\pi kT}\right)^{1/2} \exp\left(-m^* V_x^2/2kT\right) dV_x$$

For electron flow from the metal to semiconductor, the barrier height remains the same as without bias, this current therefore can be obtained from equation (5.3.5) with $V_a = 0$ as

$$J_{ms} = -AT^2 \exp(q\Phi_{bn}/kT) \quad \cdots \quad (5.3.6)$$

From equations (5.3.5) and (5.3.6) the Schottky barrier emission model current density is given by

$$J = J_{se}\left[\exp(qV_a/kT) - 1\right] \quad \cdots \quad (5.3.7)$$

where $J_{se} = AT^2 \exp(q\Phi_{bn}/kT)$ ··· (5.3.8)

5.3.2. Diffusion model

The assumptions in this model are (1) the barrier thickness is much larger than the carrier mean free path so that electrons do collide within the depletion region, (2) the barrier height is much larger than kT and that (3) the boundary concentrations are independent of current flow. Under steady state conditions the current being independent of x one can use an integrating factor $\exp(-q\psi/kT)$, where ψ refers to the conduction band potential, to integrate the current flow equation

$$J = q\left[-n(x)\mu\frac{\delta\psi}{\delta x} + D_n\frac{\delta n}{\delta x}\right] \quad \cdots \quad (5.3.9)$$

$$J\int_o^W \exp\left(-\frac{q\psi}{kT}\right) = qD_n\left\{n(x)\exp-\frac{q\psi}{kT}\right\}_o^W \quad \cdots \quad (5.3.10)$$

The boundary values may be writen rs

$q\psi(o) = -q(V_b + V_n) = -q\Phi_{bn}$ where $n(o) = n_o = N_c\exp(-q\Phi_{bn}/kT)$ ··· (5.3.11)

$q\psi(W) = -q(V_n + V_a)$ and $n(W) = N_c\exp(qV_n/kT)$

substituting (5.3.11) in (5.3.10) yields

$$J = qN_cD_n\left[\exp\frac{qV_a}{kT} - 1\right] \Big/ \int_o^W \exp\left[-\frac{q\psi(x)}{kT}\right]dx \quad \cdots \quad (5.3.12)$$

The potential $\psi(x)$ can be written using step junction approximation as

$$\psi(x) = \frac{qN_d}{\varepsilon}(Wx - x^2/2) - \Phi_{bn} \quad \cdots \quad (5.3.13)$$

From equations (5.3.12) with (5.3.13) and writing for W

$$W = \sqrt{\frac{2\varepsilon(V_b - V_a)}{qN_d}}$$

we have

$$J = \frac{N_c q^2 D_n}{kT}\left[2q\frac{(V_b - V_a)}{\varepsilon}N_d\right]^{1/2} \exp\left(\frac{-q\Phi_{bn}}{kT}\right) F(V) \quad \cdots \quad (5.3.14)$$

where $F(V) = \left[\exp\frac{(qV_a)}{kT} - 1\right] \Big/ \left[1 - \exp\left\{-\frac{2q\,(V_b - V_a)}{kT}\right\}\right]$

$$\cdots \quad (5.3.15)$$

Since $qV_b \gg kT$ has been assumed for reverse bias (V_a negative) and for small forward bias (V_a positive) the exponential in the denominator of equation (5.3.15) drops out leaving for J_{dif}.

$$J = J_{sd}\left[\exp\frac{(qV_a)}{kT} - 1\right] \quad \cdots \quad (5.3.16)$$

where $$J_{sd} = \frac{N_c q^2 D_n}{kT}\left[\frac{2q(V_b - V_a)N_d}{\varepsilon}\right]^{1/2} \exp\left(\frac{-q\Phi_{bn}}{kT}\right) \quad \cdots \quad (5.3.17)$$

Though the expressions (5.3.7) and (5.3.16) are very similar to the saturation current J_{sd} of the diffusion theory varies with voltage faster but with temperature slower than the Thermionic emission theory saturation current J_{se} of equation (5.3.7).

A synthesis of the thermionic emission and diffusion theories has been made by Crowell and Sze (ref. 1). In the nature of the Schottky barrier there having a potential minimum near the metal contact ; the electrons are carried due to bias by diffusion to this point ; wherefrom the migration to the metal, is described by a thermionic recombination velocity V_R at the metal semiconductor interface. This enables incorporation of electron phonon scattering and quantum mechanical reflection at the metal semiconductor interface. It is shown that at low fields as long the metal acts as a perfect sink for electrons crossing the potential minimum, the thermionic emission theory is applicable. At higher fields quantum mechanical reflections and quantum tunnelling caused increased recombination changes the regime of operation to thermionic field emission.

It has been shown that with $V_R = (AT^2/qN_c)$, A being the effective Richardson constant the current density can be written as :

$$J = \frac{qN_cV_R}{1+\frac{V_R}{V_d}} \exp\left(-q\Phi_{bn}/kT\right) \left[\exp\left(-qV_a/kT\right)-1\right] \quad \cdots \quad (5.3.18)$$

$$\text{where } V_d = \left[\int_{x_m}^{W} \frac{q}{\mu kT} \exp \frac{-q}{kT} \left(\Phi_{bn}+\psi\right) dx\right]^{-1} \quad \cdots \quad (5.3.19)$$

is the effective diffusion velocity between W to x_m (cf. fig. 5.3.1). If the electron distribution between x_m to W is Maxwellian and no incident electrons except the original equilibrium ones get reflected from the metal then the semiconductor acts as a thermionic emitter. If $V_R \leqslant V_d$ then thermionic emission mostly governs the current flow on the other hand if $V_R \geqslant V_d$ diffusion is the dominant mechanism. In this case if μ is independent of electric field the diffusion velocity will be nearly equal to the drift velocity in the semiconductor at the boundary *i.e.*, $V_d = \mu E$ and equation (5.3.18) reduces to

$$J = qN_c\,\mu E \exp\left(-q\Phi_{bn}/kT\right)\left[\exp\left(\frac{qV_a}{kT}\right)-1\right] \quad \cdots \quad (5.3.20)$$

with increased field the probability of backward scattering of electrons due to optical phonon scattering increases, that reduces the net current. In addition the carrier energy distribution deviates from Maxwellian distribution because of quantum mechanical reflection of electrons as waves and their tunnelling through the barrier. When the field enhanced tunnelling becomes predominant a rapid transition from thermionic to thermionic field emission takes place.

Taking the back scattering and field enhanced tunnelling into account the current in Ge, Si and GaAs can be written as

$$J = J_s\left[\exp\left(qV_a/kT\right)-1\right] \quad \cdots \quad (5.3.21)$$

$$\text{where} \quad J_s = A^* T^2 \exp\left(-q\Phi_{bn}/kT\right) \quad \cdots \quad (5.3.22)$$

A^* is the effective Richardson constant including the effects of optical phonon scattering and thermionic field emission. It remains more or less constant at room temperature ; in the electric field range 10^4—10^5 V/cm in most Ge, Si and GaAs Schottky diodes and the current flow is due to thermionic emission mechanism in these diodes. In Si for electrons its value being 110 A/cm^2/°K^2 while for holes it is about 30A/cm^2/°K^2.

5.3.3 Minority Injection

Under the low injection conditions we see that the Schottky barrier diode is a majority carrier diode. The minority carrier density at the bulk at W (cf. fig. 5.3.1) is, using (5.3.21) and (3.3.2).

$$p_n(\mathrm{W}) = \bar{p}_{no}\,[\exp(q\mathrm{V}_a/k\mathrm{T}) - 1] = \frac{n_i^2}{\mathrm{N}_d}\frac{\mathrm{J}}{\mathrm{J}_s} \qquad \cdots \quad (5.3.23)$$

For low bias voltage J is low, the minority injection ; hence current is also low particularly because the drift and diffusion components nearly balance out. But for high bias voltages the drift field increases ; hence ultimately drift current dominates and hence appreciable minority current flows. Thus the minority to majority injection ratio increases in high current cases. From (5.3.23) we note that to reduce minority injection N_d has to be high and J_s large and n_i small *i.e.*, high doping, small barrier and large band gap are necessary.

The storage time on the other hand

$$\mathrm{T}_s = \frac{\mathrm{Q}_{\text{stored}}}{\mathrm{J}} = \frac{\int_{\mathrm{W}}^{\mathrm{W}_o} (qpdx)}{\mathrm{J}}$$

assuming exponential minority distribution *i.e.*, large bulk

$$(\mathrm{W}_o - \mathrm{W}) \gg \mathrm{L}_p$$

$$\mathrm{T}_s = \frac{qn_i^2\mathrm{L}_r}{\mathrm{N}_d\mathrm{J}_s} \qquad \cdots \quad (5.3.24)$$

which is inversely proportional to current density and lies between $1n$ sec and $0.01n$ sec. in the 10 – 1000 amp/cm^2 range.

Taking account of the voltage dependence of A* the effective Richardson constant and the Schottky barrier lowering, the J/v characteristic can be written as

$$\mathrm{J} = \mathrm{J}_s\left[\exp\left(\frac{q\mathrm{V}_a}{nk\mathrm{T}}\right) - 1\right] \qquad \cdots \quad (5.3.25)$$

From this the diode a.c. conductance is

$$g_j = \frac{d\mathrm{I}}{d\mathrm{V}} = \frac{q\mathrm{JA}}{nk\mathrm{T}} \qquad \cdots \quad (5.3.26)$$

But the junction a.c. resistance cannot be expected to tend to zero for very large currents since there would be some amount of series bulk resistance and contact resistances. If the total bulk plus junction forward a.c. resistance is r_s and total capacitance at forward (~ 0.1v) bias situation c_f then the diode cut off frequency (f_o) can be written as

$$f_o = \frac{1}{2n r_s c_f} \qquad \cdots \quad (5.3.27)$$

Because at forward bias $r_j = (1/g_j)$ decreases faster than c_f increases, the small forward bias f_o is used as a practical figure of merit. This high cut-off frequency arises mainly from the high mobility of *n*-type carriers.

As for the reverse characteristics, the Schottky barrier lowering being field dependent, the reverse current increases with applied voltage, for example, over 0.01v to 20v from 0.1*n*A to 10*n*A for typical diodes before the junction avalanche breakdown voltage is reached. This implies reverse resistance decreases but for typical 0.5 micron diameter diodes $10^{10} - 10^{12}$ ohms is typical of Au-Si and Au-GaAs diodes, the silicon being lower. The forward resistance on the other hand, is higher for silicon ~ 10 ohms while for GaAs ~ 1 ohm.

5.4 Point Contact diode

It is made by making a sharp point contact between a semiconductor and a fine metal wire. The contact may be just a mechanical contact or may be formed by passing a controlled electric discharge which may not or may even form a *p-n* junction of very small area. This diode used as microwave rectifiers usually has larger forward a.c. resistance than the Schottky diodes mentioned above and this is because of smallness of area, while the reverse current increases faster than the Schottky diode ; also the reverse resistance is considerably low and changes with bias viz., say, $10^5 - 10^8$ ohms from 0.2v – 4v voltage change before avalanche breakdown sets in. This is because of large field concentration beneath the point contact. In this rectifier I/V characteristics cannot be predicted theoretically in general. This is because the characteristics depend on many parameters, such as, contact area, contact pressure, whisker composition, heat and time duration of forming discharge, surface treatment before and after contact formation etc. Because of its small area, capacitance is low which ultimately makes it useful in the MW range even if the a.c. forward resistance is higher.

5.5 Space charge limited diode

If a diode is made with an ohmic contact at one end and a rectifying contact at the other (with appropriate metals at the ends) from a semiconductor either intrinsic Si or little or with no doping of higherband gap material then we would expect charge relaxation time of the semiconductor medium to be appreciable and space-charge-limited current flow to occur through the semiconductor medium. This sort of diode is called space-charge-limited (S.C.L.) diode. As shown in Fig. (5.5.1) the electrons from the metal would be available in the conduction band at the ohmic contact because of its nearness to the metal F.L. (cf. Fig. 5.1.2b). Similarly, the anode would be a rectifying contact that will drain these electrons (cf. Fig. 5.1.1b). As for the current flow we can write

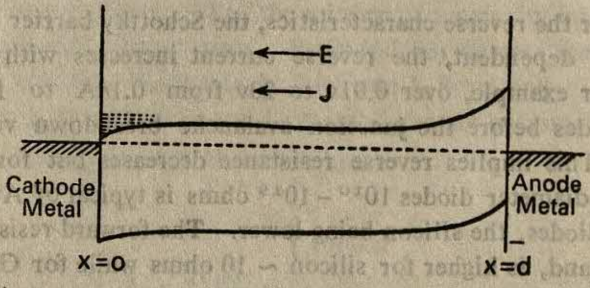


Fig. 5.5.1. Energy band diagram of S.C.L. diode formed of suitable metal contacts on appropriate semiconductor such as I*n* Cds. A*g*

$$-\mathrm{J} = \mu q n \mathrm{E} + \mathrm{D}_n q \cdot \frac{dn}{dx} \quad \cdots \quad (5.5.1)$$

and for the space charge

$$\left(\frac{d\mathrm{E}}{dx}\right) = -nq/\varepsilon\varepsilon_o \quad \cdots \quad (5.5.2)$$

So long the applied voltage is small the diffusion current is important since

$$\mu q n \mathrm{E} \leqslant \mathrm{D}_n q \frac{dn}{dx}$$

implies

$$\frac{q}{k\mathrm{T}} \int \mathrm{E}\, dx \leqslant \int \frac{dn}{n} \quad \cdots \quad (5.5.3)$$

Hence for very low currents when the applied voltage is much less than (kT/q) we have from (5.5.1) neglecting the field component and using equation (5.5.2) we have

$$J = D_n \varepsilon \varepsilon_0 \frac{d^2 E}{dx^2} \qquad \cdots \quad (5.5.4)$$

which can be thrice integrated to show that the current voltage relation will be ohmic in this voltage range.

When the applied voltage increases and the field current dominates we have from (5.5.1) neglecting the diffusion contribution to current and using equation (5.5.2),

$$J = \mu \varepsilon \varepsilon_0 E \frac{dE}{dx} \qquad \cdots \quad (5.5.5)$$

Integrating equation (5.5.5) and assuming at the cathode the potential minimum has reached *i.e* $E = 0$ at $x = 0$, eliminates the integration constant and gives

$$\mu \varepsilon \varepsilon_0 \frac{E^2}{2} = Jx \qquad \cdots \quad (5.5.6)$$

From equation (5.5.6) we have

$$E = \sqrt{\frac{2J}{\mu \varepsilon \varepsilon_0}}\, \sqrt{x}\, . = -\frac{dV}{dx}, \qquad \cdots \quad (5.5.7)$$

hence integrating from cathode to anode, squaring and rearranging

$$J = \frac{9}{8} \frac{\varepsilon \varepsilon_0 \mu V^2}{d^3} \qquad \cdots \quad (5.5.8)$$

That is the current is directly proportional to the square of voltage and inversely proportional to the cube of thickness. This is the Child's law analogue of solid state diodes, also known as the Mott-Gurney law since Mott and Garney suggested this phenomenon. This shows that the contact exponential characteristic can be swamped by bulk characteristic of Schottky diode if the bulk resistivity is very high.

Next we enquire about the speed of space charge limited diodes. The transit time of carriers through the diode can be calculated by

$$T = \int_0^d \frac{dx}{v} = \int_0^d \frac{dx}{\mu E} \qquad \cdots\cdot \quad (5.5.9)$$

substituting for E from equation (5.5.7) in (5.5.9)

$$T=\sqrt{\frac{\varepsilon\varepsilon_0}{2J\mu}}\int_0^d \frac{dx}{\sqrt{x}}=2\sqrt{\frac{\varepsilon\varepsilon_0}{2J\mu}}\,d \qquad \cdots \quad (5.5.10)$$

Substituting for J in equation (5.5.10) from (5.5.8) and rearranging

$$T=\frac{4}{3}\cdot\frac{d^2}{\mu V} \qquad \cdots \quad (5.5.11)$$

The transit time in diffusion limited *p-n* junction diodes was found (cf. equation 4.14.24) to be $T_B=\frac{W^2}{2D_P}$, hence comparing the transit times we find

$$(T_{B/T_{s.c.l.}})=\frac{W^2}{2D_p}\cdot\frac{3}{4}\,\frac{\mu V}{d^2} \qquad \cdots \quad (5.5.12)$$

hence for the same diode dimensions and same diode body and 2-3 volts applied to the S.C.L. diode gives $\frac{3}{8}\frac{qV}{kT}\sim 40$ times greater transit times in *p-n* diffusion limited diode. Hence the S.C.L. diode is very much faster in operation than the *p-n* junction diode,

The reverse characteristic of S.C.L. diode is quite good and comparable to the GaAs Schottky diode.

5.6 Varactor diode

The word varactor is derived from variable reactor. This is basically either a *p-n* junction or metal semiconductor junction diode whose capacitance variation with voltage under reverse bias is the source of this reactance variation. In some special varactors hyper abrupt junction or hyper doping in the semiconductor of a *p-n* jnnction diode is utilised. In either case, one side has or acts as if has very heavy doping; only the depletion region of the semiconductor comes in the capacitance variation. To take a special case, say, we need a varactor which will be used to tune an inductive circuit such that the resonant frequency is varied linearly with voltage, what is the doping profile needed for the semiconductor region ? Since the junction is one sided, we assume a donor doping of the form bx^m and write the Poission equation for the semiconductor as

$$\frac{d^2V}{dx^2}=-\frac{bqx^m}{\varepsilon\varepsilon_0} \qquad \cdots \quad (5.6.1)$$

Integrating once and evaluating the integration constant by the condition that the field is zero at the depletion region boundary at $x=d$ we have

$$\frac{dV}{dx}=-\frac{b.q}{\varepsilon\varepsilon_0(m+1)}(x^{m+1}-d^{m+1}) \quad \dots \quad (5.6.2)$$

Integrating once more with the condition $V=0$ at $x=0$ gives

$$V=-\frac{bq}{\varepsilon\varepsilon_0(m+1)}\left(\frac{x^{m+2}}{m+2}-d^{m+1}.x\right) \quad \dots \quad (5.6.3)$$

Hence the voltage across the depletion region V_t is

$$V_t=\frac{bqd^{m+2}}{\varepsilon\varepsilon_0(m+2)} \quad \dots \quad (5.6.4)$$

The student can easily show that the depletion region charge is $Q=\int_0^d qbx^m dx=\frac{qbd^{m+1}}{m+1}$

with it using equation (5.6.4)

$$C_j=\frac{dQ}{dV}=\frac{\varepsilon\varepsilon_0}{d} \quad \dots \quad (5.6.5)$$

or

$$C_j=\varepsilon\varepsilon_0\left[\frac{bq}{(m+2)\varepsilon\varepsilon_0 V_t}\right]^{1/m+2} \quad \dots \quad (5.6.6)$$

For an LC circuit $f=\frac{1}{2\pi\sqrt{LC}}$ or $C\alpha\frac{1}{f^2}$ ... (5.6.7)

Equating (5.6.6) and (5.6.7) we note that in order that V could be proportional to the frequency f, one has to have

$$\frac{1}{m+2}=2 \quad \text{or} \quad m=-\frac{3}{2} \quad \dots \quad (5.6.8)$$

Thus we see that a doping of the form $bx^{-3/2}$ would nearly satisfy our requirement since we have ignored the header capacitance C_h. Such a doping profile is known as hyper doping profile.

The equivalent circuit of the varactor diode is shown in Fig. (5.6.1) where $C=C_j+C_h$, R_s is the series resistance due to the bulk semiconductor and R_p is the parallel equivalent resistance for generation, diffusion and leakage currents. Both C_j hence C and R_s

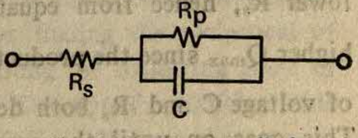


Fig. 5.6.1. Varactor diode equivalent circuit

decrease with increase of reverse voltage while R_p generally increases with voltage when the voltage is below a certain limit. The efficiency of a varactor is expressed by its Q which expresses its energy storage capability in relation to its loss and can be easily shown to be

$$Q=\frac{1}{\frac{1}{\omega CR_p}+\omega CR_s} \quad \cdots \quad (5.6.9)$$

assuming $R_p+R_s \simeq R_p$.

Equation (5.6.9) shows that the maximum Q occurs at radian frequency

$$\omega_{qm}=\frac{1}{C\sqrt{R_pR_s}} \quad \cdots \quad (5.6.10)$$

and

$$Q_{\max}=\tfrac{1}{2}\sqrt{R_p/R_s} \quad \cdots \quad (5.6.11)$$

For a certain applied voltage equation (5.6.9) shows that at low frequencies compared to ω_{qm}, given by equation (5.6.10), the value of $Q\rightarrow\omega CR_p$ and for high frequencies $Q\rightarrow(1/\omega CR_s)$. Therefore, the variation of Q can be depicted as shown in Fig. (5.6.2). For the same frequency a higher voltage will give lower C, higher R_p lower R_s, hence from equations (5.6.10) and (5.6.11) higher ω_{qm} higher $Q_{\max}$ since the product $(\omega_{qm}Q_{\max})=\frac{1}{2CR_s}$ while with increase of voltage C and R_s both decrease hence (ω_{qm} and $Q_{\max}$) increase. This goes on until the reverse voltage approaches the avalanche breakdown of the junction, when this increase is expected to stop.

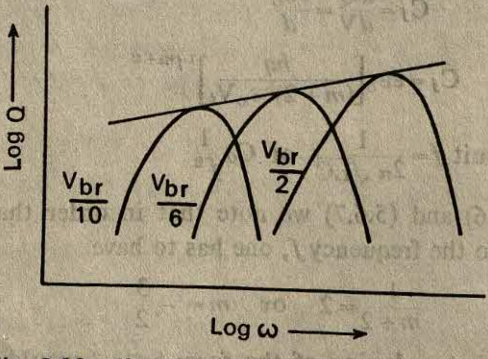


Fig. 5.6.2. Varactor diode quality factor variation with frequency as fraction of breakdown voltage as the parameter

References

1. *Physics of semiconductor devices:* S. M. Sze. Wiley Eastern limited (1979).

2. *Solid State Physical Electronics*: A. Vander Ziel. Prentice-Hall of India, Pvt. Ltd. (1971).

Problems

5.1. A metal-semiconductor diode of 10 micron diameter has a capacitance of 0.02 P*f* at 10 volts reverse bias. Calculate (*i*) the barrier layer thickness, (*ii*) the dopant concentration N_d assuming it to be constant and (*iii*) find the maximum field in the barrier.

(Assume $\varepsilon_r = 12$)

5.2. The square of the inverse of capacitance of a metal-semiconductor (Si) diode under reverse bias is seen to increase with bias voltage as $(1/c^2) = (100 + 300V)10^{24}$ where V is in volts and *c* is in p.f. Calculate the built in potential of the depletion regions and the uniform dopant concentration assuming a diode diameter of 20 microns. Find the depletion region width at zero bias and the Si doping level assuming it to be uniform.

5.3. How would you measure the barrier height of a Schottky barrier diode by means other than reverse bias capacitance measurement ?

Hint : (a) Find current for a fixed forward voltage and observe the forward current J for various temperatures. Plot log J versus reciprocal of *k*T (cf. eqns. 5.3.12 and 5.3.13). Find the slope.

(b) The short circuit photo current is proportional to $(h\nu - q\phi_{bn})^2$ so long $(h\nu - q\phi_{bn}) \gg kT$. Plot sequare root of this current against $h\nu$ and extrapolate.

5.4. For a metal semiconductor junction or an one side heavy doped *p-n* junction work out a method for finding the reciprocal of the doping profile versus depth *x* along the following lines of argument. Write the capacitor as $C = dQ/dV$. Express this more accurately as $dV = dQ(x + \delta x/2)/A\varepsilon$. If on a d.c. reverse bias a constant current $I_m \cos \omega t$ causes the charge change then find $dQ = (I_m \sin \omega t/\omega) - qN_a A\, \delta X$. From these two equations find the voltages set up across the capacitor apart from a d.c. (Ref. J. A. Copeland, IEEE Trans. ED May 1969, p. 445).

Ans. Fundamental $(I_m X/\omega A \varepsilon)$

Second harmic $(I_m{}^2)/4\omega^2 q N_a A^2 \varepsilon$.

5.5. What is a Mott barrier ?

5.6. A Schottky barrier diode of 0.01 cm^2 area is made on n^+ substrate with n-epilayer doping of 10^{15}/c.c. The n^+ subtrate of doping 10^{18}/c.c. has a thickness 0.02 cm, and has a lifetime of 10^{-8} secs. The n-doped epilayer has 10 micron thickness and a lifetime of 10^{-6} secs. For a current of 250 mA find the stored hole charge in pico-coulombs in the epilayer at 300°K. If the current is increased by a factor of 20, how does the minority charge storage change ?

5.7. In the energy band diagram of a Schottky barrier with deep donors in the semiconductor the Fermi level cuts through the deep donor level. The depth of this crossing from the met-sem interface increases with reverse bias. Show that if a reverse bias Step voltage δV is applied on a reverse bias of V_0 the capacitance should have changed from C_0 to C_1 if all the donors were instantaneously responding,

where $C_0{}^2 = q\, \varepsilon\, A^2 N_{10}/2(V_b + V_0)$ and

$C_1{}^2 = q\, \varepsilon\, A^2 N_1/2(V_b + V_1)$ with $N_1 = N_{10} + \delta N_1$

and $V_1 = V_0 + \delta V$, δV being the fast voltage step.

5.8. In problem (5.7) above argue to establish that C_1 is a function of time even if the pulse has ceased to exist since the deep energy levels are not as fast and that if there are multiple deep levels

$$\delta N_1(t) = \sum_i \delta N_1{}^i(O) \exp - (e_n{}^i + e_p{}^i)t$$

where $e_n{}^i$ and $e_p{}^i$ are the emission rates of electrons and holes at the i-th deep level. Note that, in general, $\delta N_1{}^i(O)$ itself is a function of $e_n{}^i$ and $e_p{}^i$ also.

5.9. From the expressions in problems (5.7) and (5.8) above and that

$$f(b) = c^2(t) - C_0^2 = \Sigma \, B^i \exp(-t/T^i)$$

where $B^i = q\varepsilon A^2 \, \delta N_1^i(O)/2(V_b + V_o)$ and $T^i = (e_n^i + e_p^i)^{-1}$.

5.10. From the results of problem (5.9) above show that

$$t\frac{df(t)}{dt} = \sum_i B^i(-t/T^i)\exp(-t/T^i)$$

and argue to establish that $t(df/dt)$ has a maximum value of (B_i/e) at $t = T^i$. Note that this is the basis of one technique by which deep level's energy band location and trap characteristics can be found out. (cf. Japanese Journal of Applied Physics, Vol. 19, No. 6, June 1980, pL 335).

Chapter 6

JUNCTION TRANSISTOR

In this section we discuss the d.c. characteristics of Junction Transistors.

6.1 Working Principle

A junction transistor can be either a *p-n-p* or an *n-p-n*, a two junction three electrode device as shown in fig. (6.1.1). One of the

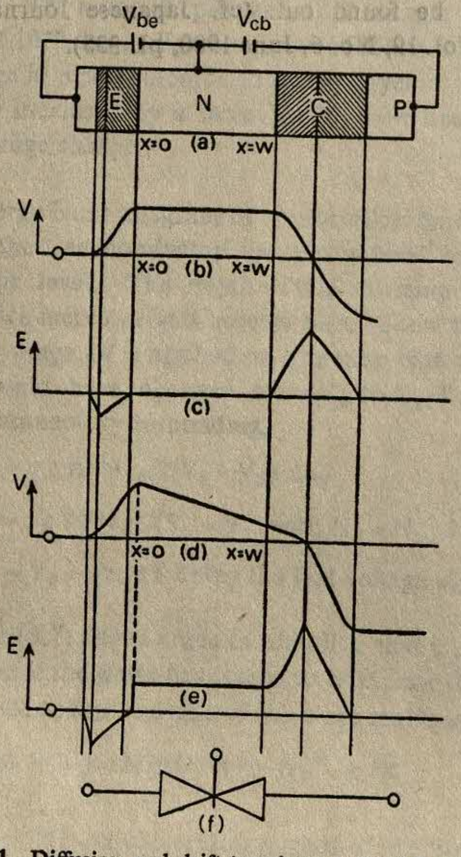


Fig. 6.1.1 Diffusion and drift transistor potential variation figs. (b) & (d) also field variation figs (c) & (e) respectively

junctions is called the emitter (E) junction which is normally forward biased and the other is called the collector (C) junction that is normally reverse biased. The whole structure is a single crystal with metallic electrodes from ohmic contacts on the three portions of the semiconductor bulk.

By forward biasing the emitter junction minority carriers are injected at the left hand boundary ($x=0$) of the base region (holes in *n*-region of *p-n-p* and electrons in the *p*-region of *n-p-n*). These injected minority carriers travel through the base region to the collector junction. Those surviving recombination reach the collector junction boundary ($x=W$) and get drained off to the collector electrode. Note the direction of the field in the collector depletion region causing this drainage. The minority carriers those recombine, anihilate equal number of majority carriers which are constantly supplied by the base terminals as base current. Depending on the doping profile in the base region the junction transistor are divided into two types the diffusion and the drift type. In the diffusion transistor base doping is uniform while in the drift type the base doping is non-uniform (exponential in our case) Diffusion and drift transistor junction and base region potential variations are shown in Figs. (6.1.1*b* & *d*). In diffusion transistor the minority carriers travel due to diffusion only while in the drift transistor they travel due to diffusion and the drift caused by the built in field. The minority carriers find an accelerating field, in the collector depletion region in both the types of transistors, while in the drift transistor, the accelerating field exists in the base body ($x=0$ to $x=W$) also.

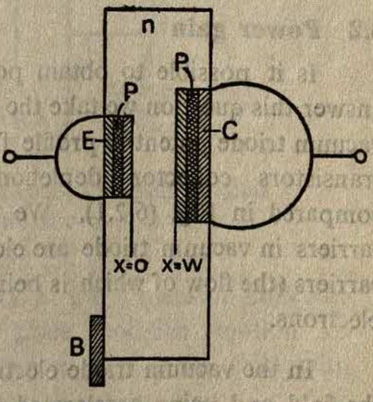


Fig. 6.1.2 *p-n-p* alloyed transistor

For a better understanding of the device operation we show in Fig. (6.1.2) a *p-n-p* alloyed transistor. In this figure the base is N-type semiconductor and emitter (E) and collector (C) bulks are p-type with alloyed metal contacts. The base B is an ohmic alloyed contact. Note that the base region is very thin. It is deliberately made very thin (2 μm to 0.5 μm to have the base thickness $W \ll L$ the minority carrier diffusion length) so that the recombination loss of carriers in transit through the base region ($x=0$ to $x=W$) is small and most of the minority carriers injected (at $x=0$) reach the collector boundary ($x=W$). The base contact on the other hand is far away from the emitter ($d \geq 5L$) such that the base terminal only supplies the recombination replenishing majority carriers. Emitter to collector potential profile combined with the nearness ($\sim$ 1 micron) of emitter and collector causes the transit of most of the minority carriers to the collector.

The word transistor has been derived from 'transfer resistor'. Because the base region is very thin the recombination loss of injected carriers are small. Hence whatever minority current is injected nearly all of it flows to the collector. The current which crossed the forward biased small resistance region gets transferred to the reverse biased high resistance region. Current remaining same the resistance is (changed) transferred, hence the name.

6.2 Power gain

Is it possible to obtain power gain from such a device ? To answer this question we take the example of the vacuum triode. The vacuum triode potential profile from cathode to anode and the *n-p-n* transistors collector depletion region potential variation are compared in Fig. (6.2.1). We take an *n-p-n* transistor because the carriers in vacuum triode are electrons while in N-P-N the minority carriers (the flow of which is being controlled in the P-type base) are electrons.

In the vacuum triode electrons flow opposite to the direction of the field and being accelerated gains energy from the field. Such a steady stream of electrons flowing means a steady current flowing and energy being extracted from the supply. A small change (say increase) of current produced by the grid control causes an extra potential drop in the load resistor R connected to the anode. The tube drop decreases. Hence the potential profile changes to the one dotted causing the field to change (decrease) by $\triangle E$. The electrons representing the current has to move being retarded by (giving energy to) the field $\triangle$E as shown in Fig. (6.2.1). That means the tube dissipation falls short of this incremental energy. This energy was obtained by the electrons from the supply but the incremental energy by which the tube dissipation has fallen short of is lost in the load resistor. This is the mechanism of power conversion in a vacuum triode.

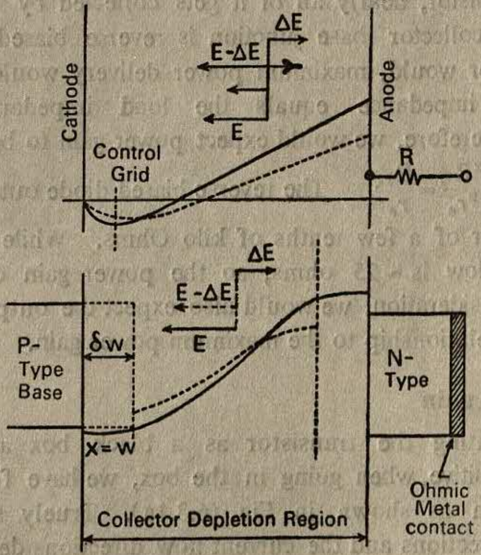


Fig. 6.2.1. Comparison of V_{AC} tube & collector depletion region potential variations during incremental current changes

Exactly similar phenomena occurs in the transistor collector region. Scattering phenomena only gives modified motion of individual carriers but they move en masse due to the field with velocity depending on the field. In this depletion region carriers get accelerated by the existing field E at each point. Also the incremental carriers get opposed by the incremental field at each point. Thus, even if the semiconductor body is entirely passive, power gain is obtainable as in a conventional triode. In *p-n-p* transistors the carrier charge being of opposite type the field directions (including the incremental part) are also opposite to that of *n-p-n* or vacuum triode. The decrease in field caused by the decrease in available voltage across the depletion region, decreases the depletion region width. This, in turn, causes the effective width W of the base region to increase (by δW, say). From the above mentioned mechanism of power gain one would, therefore, expect some sort of relationship of the power gain to the variation of base width (δW) with collector base voltage variation (δ V_{CB}).

It has been mentioned that the minority current which crosses the emitter resistor, nearly all of it gets collected by the collector. Whereas the collector base junction is reverse biased. Seen as a circuit engineer would, maximum power delivery would occur when the output impedance equals the load impedance. At low frequencies, therefore, we would expect power gain to be of the order of $G = \frac{P_{out}}{P_{in}} = \frac{i^2 R_L}{i^2 r_e} = \frac{R_L}{r_e}$. The reverse biased diode output resistance is of the order of a few tenths of kilo Ohms. While r_e for 1 mA d.c. current flow is ~ 25 ohms, so the power gain can be large. From this consideration, we would also expect the output impedance having some relationship to the maximum power gain.

6.3 Current gain

Representing the transistor as a black box and using all currents as positive, when going in the box, we have for the output short situation as shown in Fig. (6.3.1a). Truely speaking, the carrier flow directions and the current flow directions depend entirely on the type of minority carrier concerned.

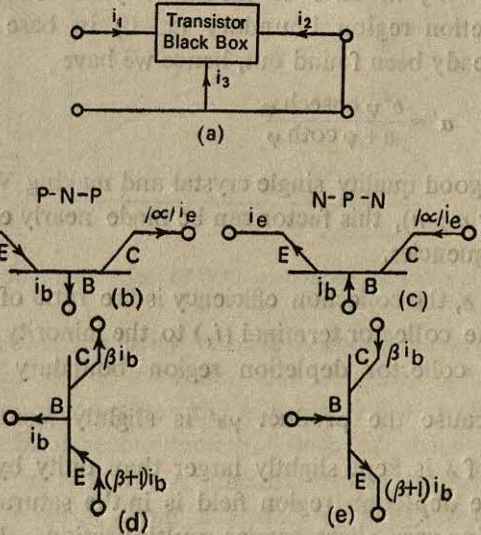
The short circuit current gain is defined in general as the ratio of output current to input current for no output voltage variation (short curcuit for a.c.). Regarding α as a positive quantity (cf. Fig. 6.3.1a with b) common base (CB) configuration short circuit current gain

$$-\alpha = \frac{i_2}{i_1}\bigg|_{V_{CB\ \text{const.}}} = \frac{i_c}{i_e}\bigg|_{V_{CB\ \text{const.}}} \quad \cdots \quad (6.3.1)$$

while for common emitter configuration

$$\beta = \frac{i_2}{i_1}\bigg|_{V_{CE\ \text{const.}}} = \frac{i_c}{i_b}\bigg|_{V_{CE\ \text{const.}}} \quad \cdots \quad (6.3.2)$$

Note that α defined as $\left(\frac{i_2}{i_1}\right)$ in Fig. 6.3.1 is negative while β is positive because from physics of the transistor for p-n-p transistor, collector current is going *out* but emitter current is going *in* and α is their ratio while collector current is going *out* and base current is also going *out* ; hence β, their ratio is positive.



6.3.1 (a) Depicts the circuit direction of currents ; (b) & (c) current direction as given by physics of transistor C.B. ; (d) & (e) current directions as given by physics for C.E. For P.N.P & N.P.N transistors respectively

Counting through the process of injection of minority carriers, their transport and collection, the current gain α can be written as the product of 'injection efficiency' (γ), transport factor (α') and 'Collection efficiency' (δ).

$$-\alpha=\frac{i_c}{i_e}=\gamma\,\alpha'\,\delta \qquad \cdots \quad (6.3.3)$$

for a *p-n-p* transistor, $-\alpha=\left(\frac{i_{pe}}{i_{ne}+i_{pe}}\right)\left(\frac{i_{pc}}{i_{pe}}\right)\left(\frac{i_c}{i_{pc}}\right)$

where for a *p-n-p* transistor for example, γ is ratio of hole current injected at base by the emitter to the total emitter current, that is

$$\gamma=\frac{i_{pe}}{i_{ne}+i_{pe}}=\frac{1}{1+\frac{i_{ne}}{i_{pe}}} \qquad \cdots \quad (6.3.4)$$

We have seen that by doping the emitter heavily compared to the base doping it is possible to make γ nearly equal to unity. i_{ne} being very much less than i_{pe}, we do not speak of its phase in small signal theory, *i.e.*, $\gamma = 1 \angle 0°$. The transport factor α' is defined as the ratio of minority current reaching the collector depletion region boundary ($x = W$) in base to the minority current injected at the emitter depletion region boundary ($x = 0$) in base material. This ratio has already been found out, hence we have

$$\alpha' = \frac{e^{\theta}\,\psi\,\text{cosech}\,\psi}{\theta + \psi\coth\psi} \qquad \cdots \quad 4.11.1$$

Using good quality single crystal and making $W \leqslant L_p$ for *p-n-p* ($W \leqslant L_n$ for *n-p-n*), this factor can be made nearly equal to unity for very low frequencies.

Lastly δ, the collection efficiency is the ratio of the total current flowing in the collector terminal (i_c) to the minority current (i_{pc}) that reached the collector depletion region boundary ($x = W$), that is $\delta = \frac{i_c}{i_{pc}}$. Because the product $\gamma\alpha'$ is slightly less than unity, the magnitude of δ is kept slightly larger than unity by proper biasing, such that the depletion region field is in the saturation velocity (v_s) region, causing very slight carrier multiplication. For simplicity we shall assume δ to be equal to unity *i.e.*, ($\delta = -1 \angle 0°$), which implies that we are neglecting the transit time of the carriers through the collector depletion region width (d). This time can be estimated from $T_c = \frac{d}{v_s}$.

We see that provided $\gamma \simeq 1$ and $|\delta| = 1$ the CB current gain α is solely dependent on the transport factor α', the low frequency value of which we call α_0.

It is seen that α_0 the low frequency current gain depends largely upon current magnitude. Our aim is to inject minority carriers and in that we have neglected the recombination in the emitter depletion region. At any place the total current remains constant while recombination of minority current (holes in *p-n-p*) implies decrease of minority current (i_{pe}) with consequent increase of majority current (i_{ne}). With increase of forward bias the emitter depletion region width decreases. On that count recombination in this depletion region decreases causing the majority current (i_{ne}) to decrease. This improves γ with increased bias as shown in Fig. (6.3.2) curve 'a'.

With increased minority injection level at the base both the minority, as well as majority concentration increases. This majority increase in base causes greater minority current injection in emitter bulk (increasing i_{ne}) causing γ to decrease. On the other hand, majority concentration increase decreases the minority life time in the base causing increased base recombination. Both of these two causes deteriorate α_0. Moreover, a *p-n* junction means this is a surface different in some respects from a perfect lattice and this causes increased surface recombination with increased current. The deterioration of α_0 caused by these three reasons are shown in Fig. (6.3.2) curve 'b'. Thus, we see that there are two competing processes which determine the variation of the value of α_0 with current, the result being α_0 though less than unity has a maximum at a certain emitter current as shown in Fig. (6.3.2) curve 'c'.

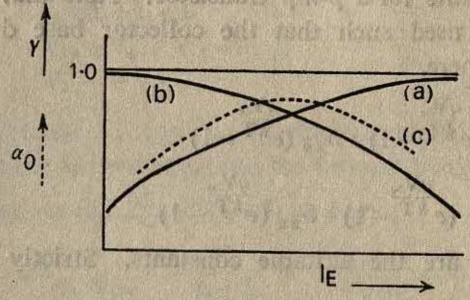


Fig. 6.3.2. Shows the effect of D.C. current on the value of α_0.

6.4. D.C. Characteristics

The common base biasing circuit with circuit notation current directions are shown in Fig. (6.4.1). It is evident that basically the transistor is two *p-n* diodes connected back to back or front to front (cf. Fig. 6.1.1f), with the base region common to both the

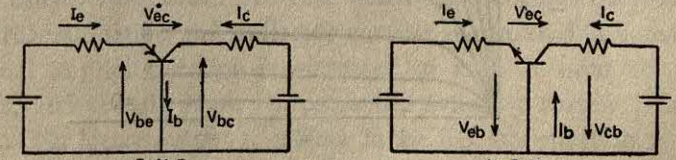


Fig. 6.4.1. Biasing polarities for the two transistor types.

diodes. With this in mind, emitter collector terminal currents can be written as the superposition of the two diode currents with proper choice of constants for a *p-n-p* transistor. Note that the collector bias polarity is used such that the collector base diode is taken forward biased here.

$$I_e = b_{11}\left(e^{\frac{qV_{be}}{kT}} - 1\right) + b_{12}\left(e^{\frac{qV_{bc}}{kT}} - 1\right) \qquad \cdots \quad (6.4.1)$$

$$I_c = -b_{21}\left(e^{\frac{qV_{be}}{kT}} - 1\right) + b_{22}\left(e^{\frac{qV_{bc}}{kT}} - 1\right) \qquad \cdots \quad (6.4.2)$$

where the *b*'s are the suitable constants. Strictly speaking, the

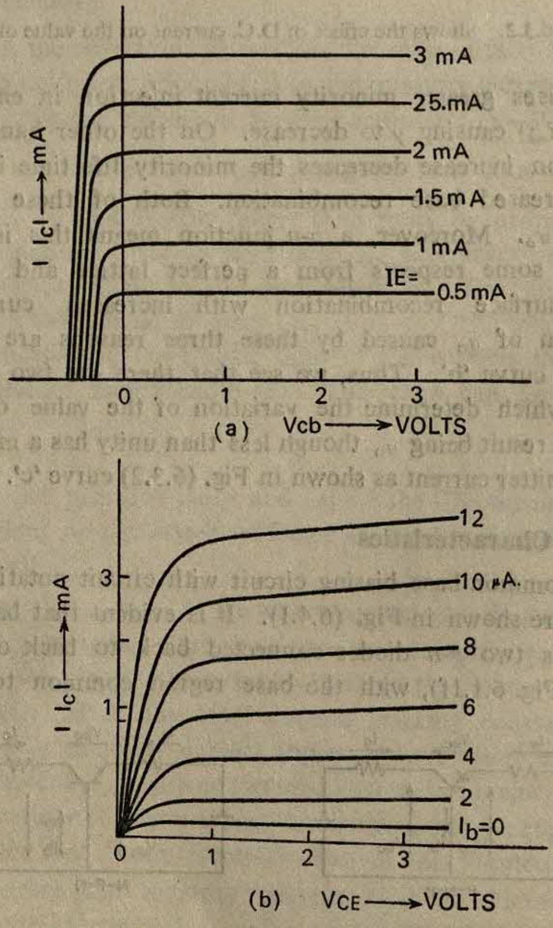


Fig. 6.4.2. (a) CB output and (b) CE output characteristics.

complete transistor behaviour should be describable by equations (6.4.1) & (6.4.2). We must also have

$$\left.\begin{array}{l} I_e + I_c + I_b = 0 \\ V_{be} + V_{ec} + V_{cb} = 0 \\ V_{bc} = -V_{cb} \end{array}\right\} \qquad \cdots \quad (6.4.3)$$

For *n-p-n* transistor V_{be} and V_{bc} would be replaced by V_{eb} and V_{cb} respectively in the current equations for forward biasing the diodes.

From equations (6.3.2) and (6.4.3) we have

$$\beta = \frac{I_c}{I_b}\bigg|_{V_{CE}} = -\frac{I_c}{I_c + I_e} \simeq \frac{\alpha_{d \cdot c}}{1 - \alpha_{d \cdot c}}$$

β for very low frequencies is $\beta_0 = (\alpha_0 / 1 - \alpha_0)$

Equation (6.4.2) with (6.4.1) indicates that for constant emitter current when V_{bc} changes from small positive value through zero to negative the collector current passes through zero to a constant value as shown in the diagram (6.4.2*a*). Due to the closeness of the emitter and collector and the potential profile discussed before even at zero collector base voltage nearly all of the emitted current reaches the collector. High reverse biased collector voltage slightly increases collector current because the base width decreases. The variation of α_0 with emitter current also gets reflected in the characteristics but α_0 being near unity it may not be noticeable in the characteristics. In the drift transistor this effect is much smaller than in diffusion transistor. The common emitter characteristics, on the other hand, are shown in Fig. (6.4.2b). The collector base junction is reverse biased hence reverse biased leakage current $I_{eo} = (\beta_0 + 1)\ I_{co} = I_c$ for $I_b = 0$ flows as shown. Hence even for zero total base current there is injection and hence collector current. Increase of base current implies more injection hence more collector current. For a certain injection from emitter, increase of V_{CE} causes base width decrease hence base current decrease due to less recombination and collector current increases, but constant base current implies injection is increased to keep it upto the previous value, and this, in turn, increases the collector current ; this is the reason for lesser output resistance in the CE configuration.

For the drift transistor, the base width variation becomes small thus increasing the output resistance. β_0 correspondingly goes through a maximum and this variation also gets reflected in the characteristics.

6.5. Punch through voltage

As the collector base reverse bias voltage increases the collector base junction depletion region width increases causing the base region width W to decrease. With gradual increase of collector base reverse bias voltage the base width decreases until the base collector depletion boundary has reached the emitter base depletion boundary when the base region width reduces to zero. In this state the transistor base region is said to be punched through, when the whole of the base region is occupied by the space charge due to ionised dopant atoms since this region is depleted of mobile carriers. The potential and space charge for a *p-n-p* drift transistor is shown in Fig. (6.5.1) for this extreme situation. In this state since the field in

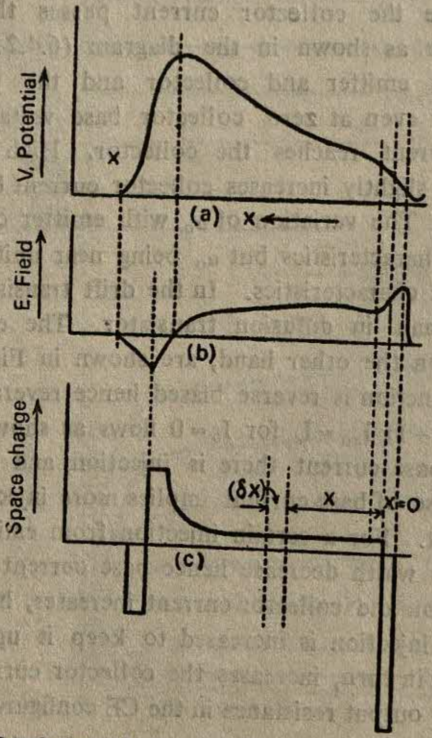


Fig. 6.5.1. Punch through state, (a) Potential, (b) Field (c) Space charge in the base region of a *p-n-p* drift transistor

the base region accelerates the injected minority carriers it is expected that a_0 the transport factor would improve. As a matter of fact, punch through type of super-β transistors having $\beta_0 \sim 1000$ or more are constructed utilizing this principle. The collector junction field in punched through state becomes high and care should be taken so that collector junction breakdown does not occur under this condition of operation. Hence we estimate the punch through voltage for exponential base doping of the transistor assuming the collector doping is high enough so that the depletion region does not extend in the collector body appreciably. Taking the origin at the collector junction, as shown in Fig. (6.5.1) if the depletion region width is to increase by δx from the value x towards the emitter a voltage δV drops across this incremental exposed space charge. This charge is given assuming unit area, by the product of the capacitance (ε/x) and the δV drop as,

$$(\varepsilon\,\delta V/x) = qN\delta x \qquad \cdots \quad (6.5.1)$$

$$\text{hence } V_{pt} = \int_0^W qN \times \delta x/\varepsilon \qquad \cdots \quad (6.5.2)$$

To be consistent with doping described in equation (4.1.12) the base doping can be written regarding x positive on left of $x = 0$ as

$$N = N_{de}.\exp\left[-2\theta\,(1 - x/W)\right] \qquad \cdots \quad (6.5.3)$$

From equation (6.5.2) with (6.5.3) we have for the collector to emitter punch through voltage

$$V_{pt} = (qN_a W^2)\,(2\theta - 1 + e^{-2\theta})/4\theta^2\varepsilon \qquad \cdots \quad (6.5.4)$$

For $\theta \to 0$ *i.e.*, for a diffusion transistor equation (6.5.4) can be reduced to give

$$V_{rt} = (qN_a W^2/2\varepsilon) \qquad \cdots \quad (6.5.5)$$

That is the punch through voltage from collector to emitter V_{pt} (C → E) is the same as V_{pt} (E → C) in a diffusion transistor. It can be shown for exponential doping of the base the ratio of these two punch through voltages for large values of θ

$$V_{pt}\,(C \to E)/V_{xt}\,(E \to C) = (2\theta - 1) \simeq 2\theta \qquad \cdots \quad (6.5.6)$$

That is the emitter to collector punch through voltage for such transistors is lower than collector to emitter punch through voltage.

This is expected since, increase of voltage in the latter case exposes less width of charge due to approach of the depletion towards increased doping as opposed to exposing large width of depletion charge due to gradual decrease of doping in the former. The emitter heavy doping required to have high injection efficiency causes the emitter breakdown to occur earlier than punch through from emitter to collector.

Normally the collector doping will not be as high as has been assumed in the above derivations for collector to emitter punch through voltage. The depletion layer will, therefore, extend in the collector side of this junction and this will need some voltage drop for its depletion. The actual collector to emitter punch through voltage is, therefore, expected to be little higher than given by the expression (6.5.4) for the drift transistor discussed above.

In order to measure the punch through voltage, a set-up as shown in Fig. (6.5.2) can be used. A variable sinusoidal voltage is used to drive to collector base junction through a diode, as shown in this Figure. The emitter and the collector of the transistor are connected to the X and Y deflecting terminals of the oscilloscope. As the a.c. voltage is increased from zero the size of the vertical trace on the oscilloscope increases. As soon as the punch through occurs, the effect of the collector voltage reaches the emitter ; hence the trace bends along the X direction. The amount of the vertical portion of the oscilloscope trace gives the corresponding punch through voltage of the transistor.

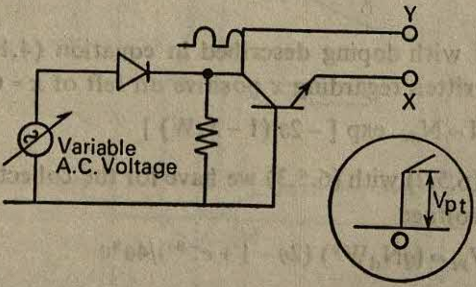


Fig. 6.5.2. Swows a simple method for the measurement of punch through voltage of a transistor

6.6. BJT breakdown characteristics

We have noted in Art. (4.17) that the junction diode breakdown is determined by the breakdown field of the semiconductor concerned and the doping levels on either side of the junction. In the junction transistor, the situation is a little different, since the junction breakdown is always determined by carrier multiplication due to collision ionisation in the junctions high field region, whereas in a transistor the collector current crossing the junction before breakdown depends on the base emitter resistance. For example, two extreme cases are shown in Fig. (6.6.1). In Fig. (6.6.1a), the transistor base is connected to the emitter while in Fig. (6.6.1b) the base is left open and in both cases collector to emitter voltage is applied. As usual the collector being negative with respect to the emitter as is necessary for *p-n-p* transistor biasing. In the first case (ignoring the small base emitter forward bias resistance) the collector junction leakage current flows from the base through the collector junction to the collector terminal. Whereas in the second case the base being open this current can not flow. But since V_{CE} is applied the emitter junction is forward biased in addition to the collector junction reverse biasing. Hence to maintain zero current out of the base terminal the base emitter forward voltage drop adjusts itself to cause minority carrier hole injection by the emitter terminal such that the base current part of it becomes the leakage current through the collector junction, thus, making the base terminal current zero. As shown in Fig. (6 6.1b), this implies the emitter injection current to be $(\beta_0+1)\,I_0 = I_0/(I-\alpha_0)$. α_0 part of this flows as collector current of the transistor while $(1-\alpha_0)$ part that is I_0 plays the role of the leakage current. Hence the collector terminal and junction current remains as $I_0\left(1+\frac{\alpha_0}{1-\alpha_0}\right)=(\beta_0+1)\,I_0$. This is the mechanism how

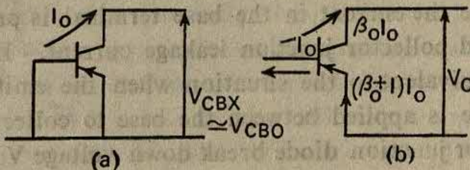


Fig. 6.6.1. Shows the extreme cases (a) Base emitter resistance zero neglecting small base medium resistance x, (b) Base emitter resistance infinite (open base).

the base terminal current is reduced to zero and can be visualised as the transistor injection caused base terminal current I_o flowing as the leakage current through the collector junction without going out to the base terminal. Thus we have in this case a current of value $(I_o/1-\alpha_o)$ entering the emitter and flowing through the collector junction as well with no current in the base terminal. This shows that the current crossing the collector junction depends on the base to emitter resistance in the intermediate cases since the base terminal current flowing through this resistance will, in addition, determine the emitter base junction biasing voltage.

Neglecting the forward bias diode d.c. resistance for the emitter-base shorted case the current in the base terminal is practically I_o, the reverse biased collector junction leakage current. Hence this is more or less equivalent to the situation when the emitter is open while the voltage is applied between the base to collector. Therefore, the collector junction diode break down voltage V_{CBO} is nearly same as the shorted base transistor breakdown voltage V_{CEX}.

In the open base case, the total collector junction current is $(I_o/1-\alpha_o)$ which is much larger than I_o. Moreover, the α_o part of the carriers while moving through the high reverse bias field causes collision ionisation and hence the transport of carriers to the collector bulk is increased by the multiplication factor M thus making the collector current much larger, which can be written as $I_o/(1-\alpha_o M)$. Thus, we see that in the open base transistor large unrestricted current can flow when $\alpha_o M=1$. This is, therefore, the junction breakdown condition for this case.

It is to be noted that near ionisation caused breakdown, the collector junction current multiplication factor M can be written from equation (4.17.6) as

$$M=\frac{1}{\left[1-\left(\frac{V_a}{V_{br}}\right)^m\right]} \qquad \cdots \quad (6.6.1)$$

where for Si and Ge, $m=3$. We can write for the diode breakdown voltage, the open emitter collector base breakdown voltage V_{CBO}. In open base case the breakdown voltage V_{CEO} can be lower due to larger junction current and that this breakdown is governed by

$$\alpha_0 M = 1 = \frac{\alpha_0}{\left[1 - \left(\frac{V_{CEO}}{V_{CBO}}\right)^m\right]} \qquad \cdots \quad (6.6.2)$$

Equation (6.6.2) gives $V_{CEO} = \sqrt[m]{1-\alpha_0} \cdot V_{CBO}$ ··· (6.6.3)

for Si and Ge, $V_{CEO} = \sqrt[3]{1-\alpha_0}\ V_{CBO} \simeq \frac{V_{CBO}}{\sqrt[3]{\beta_0}}$ ··· (6.6.4)

This implies for a $\beta_0 = 125$ the open base breakdown voltage is a fifth of the shorted base breakdown voltage. Moreover, even in the shorted base case once the avalanche breakdown has started at a comparatively larger voltage the current passing across the junction increases where again multiplication plays its role identically as in open base case. This is because the collision ionised hole carriers in a *p-n-p* goes through the collector body to the collector terminal. But the electrons (the majority carriers) gets injected into the base. This, in turn, causes equal amount of hole injection from the emitter into the base. Thus, we expect that once breakdown has started in

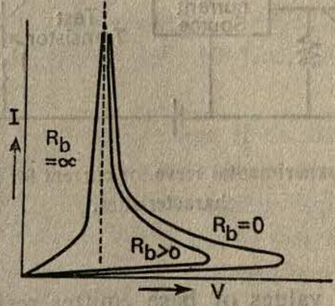


Fig. 6.6.2. Shows the transistor avalanche breakdown I/V characteristics for different base emitter resistances.

the shorted base case also, equation (6.6.2) applies after breakdown. Hence with increased current in the shorted base case the voltage across the transistor is expected to settle down to the open base breakdown voltage V_{CEO} as shown in the fig. (6.6.2).

Lastly, since present day transistors mostly have base and emitter on the same surface (planar construction) and at the edges of the emitter there is field lines crowding, breakdown at these edges may start near punch through voltage even if bulk junction breakdown is a little away when at punch through voltage a breakdown characteristic may be obtainable (cf. ref. 3).

An experimental set-up for tracing the avalanche breakdown characteristics is shown in fig. (6.6.3). A constant current source which can be a pentode tube that can supply the necessary breakdown current is connected in series with the test transistor through the supply voltage E_b.

A resistor R is connected in series as shown to limit the current through the test transistor to avoid its accidental burn out. The variable grid voltage of the tube controls the current through this transistor which can be measured by the series *m*A in the plate circuit of the tube. The floating ground d.c. V.T.V.M or a very high resistance voltmeter measures the voltage across the test transistor. Different values of base emitter resistances R_b can be connected as shown in the Fig. (6.6.3). For each R_b varying the grid voltage, noting the transistor current and the voltage across it in steps one can obtain the breakdown characteristic as shown in Fig. (6.6.2). Switching the variable supply off from the grid circuit and switching in a sawtooth voltage of proper peak value one can observe the avalanche breakdown characteristics on the oscilloscope. For this the grid to cathode voltage, being proportional to the tube current, can be displayed on the Y-axis while cathode to transistor collector voltage can be displayed on the X-axis of the oscilloscope.

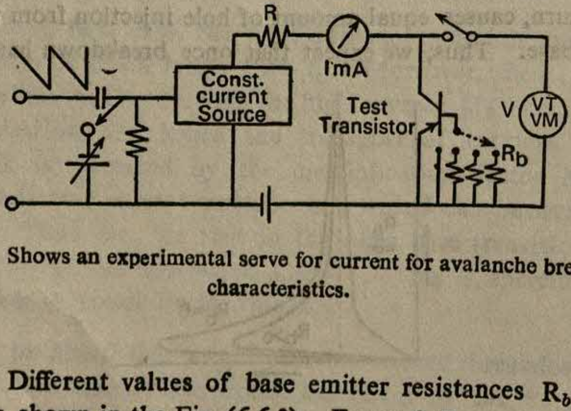


Fig. 6.6.3. Shows an experimental serve for current for avalanche breakdown characteristics.

References

1. *Fundamentals of Semiconductor Devices* : J. Lindmayer & C. Y. Wrigley, D. Van Nostrand Co. Inc. (1965).
2. *An Introduction to Semiconductor Electronics* : R. P. Nanabati, McGraw Hill Book Co. Inc., (1963)
3. *Negative resistance in a transistor under punch through conditions* : IEEE Proc. (1966), Jan. pp. 84-85. Thomas R. E., Jhonston, R. H. & Boothroyd.

Problems

6.1. Explain why β_0 versus I_c characteristics of a transistor show a maximum.

6.2. Assuming that in the normal mode of operation of a transistor $I_c = a_n I_e + I_0$ and in the inverse mode (emitter and collector interchanged) $I_e = a_i I_c + I_{eo}$ show from equation (6.4.1) and (6.4.2) that with $B = (1 - a_n a_i)$
$b_{11} = (-I_{eo}/B)$, $b_{12} = -(a_i I_o/B)$, $b_{21} = -(a_n I_{eo}/B)$ and $b_{22} = -(I_o/B)$.

6.3. Find the punch through voltage of an *n-p-n* silicon transistor having an uniform base doping 10^{15}/cc with uniform emitter and collector dopings of 10^{16}/cc. Assume $\varepsilon_r = 12$ and $W = 2$ microns.

6.4. Find the punch through voltage if the base doping is linearly decreasing from emitter junction to collector junction. Assume the emitter and the collector doping levels as 10^{16}/cc while the base doping is $b(1 - x/W) + c)$ where.$b = 10^{16}$/cc and $c = 5 \times 10^{15}$/cc. Assume an $\varepsilon_r = 12$ and base thickness 2 microns.

6.5. The base doping is exponential and is of the form $10^{16} \times \exp(-2.3 \times 10^4)$ per cc. where x is in cm, while the collector region doping level is uniform and is 10^{14}/cc over one micron followed by 10^{17} doped substrate in an *n-p-n* silicon transistor. Find for a base width of 2 microns the punch through voltage. Assume the uniform emitter doping as 10^{16}/cc.

6.6. Find the punch through voltages from emitter to collector in the three problems above and compare them with their counterparts of each problem.

6.7. Find the emitter reverse bias breakdown voltages in the three cases in the above three problems.

6.8. In problem (6.5) above if the epitaxial layer is absent what is the field at punch through in the resulting collector junction ?

6.9. Find the collector reverse bias breakdown voltage in the problems (6.3), (6.4) and (6.5) above.

6.10. In problem (6.8) above what should be the collector doping if punch through and collector breakdown are to occur for the same applied voltage with shorted emitter base ?

6.11. Draw the grid voltage versus collector to cathode voltage characteristic in the experiment of art. (6.6) when a suitably large repetitive sawtooth voltage is applied to the grid of Fig. (6.6.3) for emitter base shorted and base open cases.

CHAPTER 7

BIPOLAR TRANSISTOR A.C. OPERATION

In this chapter we shall find the equivalent circuit of the transistor applicable for a.c. operation based on the *p-n* junction diode ideas discussed in Chapter 4. In Chapter 4 it has been shown that the base region is entirely passive ; also we have obtained two admittance parameters seen from the carrier injection end, these two parameters will be utilised to derive the a.c. equivalent circuit.

7.1. Transistor A.C. equivalent circuit

It has been said that the base region of the transistor is passive entirely, the power gain of the transistor comes from the collector depletion region. Apart from the transistor action there are passive components namely the emitter to base depletion region capacitance plus the header capacitance C_e, the collector base transition region capacitance plus the header capacitance of collector and base electrodes C_c.

(*i*) With this in mind, Fig. (7.1.1*a*) shows the transistor equivalent circuit.

(*ii*) The base current, *i.e.*, the recombination replenishing majority current is supplied by the base terminal, whereas the

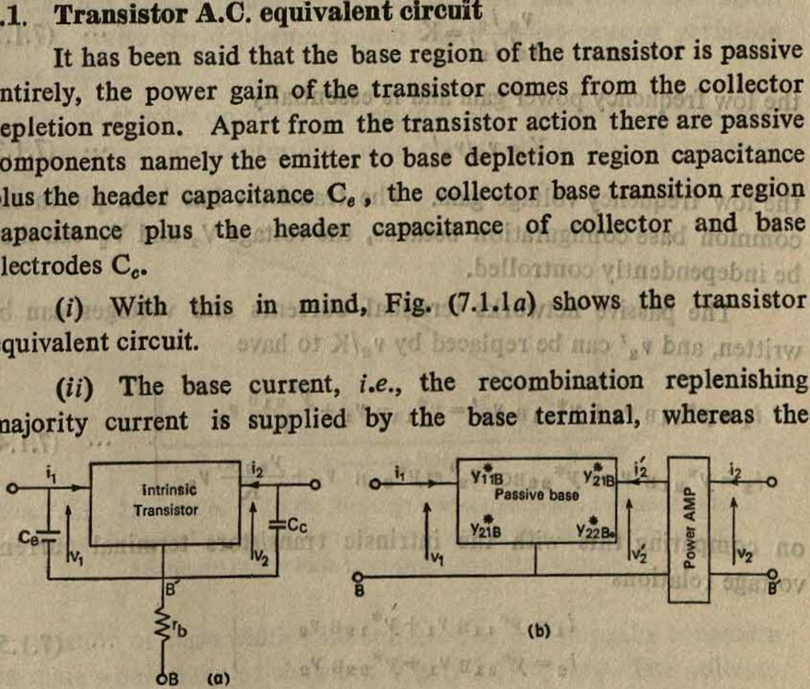


Fig. 7.1.1(*a*) Shows the transistor equivalent circuit (*b*) seperates the power gain from the passive base

recombination occurs at distributed points ; hence the distributed resistances from base terminal to these distributed points is lumped as r_b.

(*iii*) We have noted in the previous chapter that the power gain comes from the collector depletion region hence the intrinsic transistor equivalent circuit is represented in Fig. (7.1.1*b*) where the power amplifier stands for the depletion region.

We have assumed that no collector depletion region current multiplication occurs and at best there is a delay of amount $T=\frac{d}{v_s}$ hence a phase lag of ωT giving

$$i_2=i_2' \, e^{-j\omega T} \qquad \cdots \quad (7.1.1)$$

As a first approximation we neglect this delay and the complex power gain in the depletion region becomes

$$\frac{v_2\, i_2}{v_2'\, i_2'}=K \qquad \cdots \quad (7.1.2)$$

the low frequency power gain and is essentially

$$K=v_2/v_2' \qquad \cdots \quad (7.1.3)$$

the low frequency voltage gain. The transistor has been put in the common base configuration because, the voltage V_1 and V_2 can then be independently controlled.

The passive networks terminal currents and voltages can be written, and v_2' can be replaced by v_2/K to have

$$\left.\begin{aligned} i_1&=y^*_{11B}\, v_1+y^*_{21B}\, v_2'=y^*_{11B}\, v_1+\frac{y^*_{21B}}{K}\, v_2 \\ i_2&=y^*_{21B}\, v_1+y^*_{22BO}\, v_2'=y^*_{21B}\, v_1+\frac{y_{22BO}}{K}\, v_2 \end{aligned}\right\} \qquad \cdots \quad (7.1.4)$$

on comparing this with the intrinsic transistors terminal current voltage relations

$$\left.\begin{aligned} i_1&=y^*_{11B}\, v_1+y^*_{12B}\, v_2 \\ i_2&=y^*_{21B}\, v_1+y^*_{22B}\, v_2 \end{aligned}\right\} \qquad \cdots \quad (7.1.5)$$

we find that $y^*_{12B}=\frac{y^*_{21B}}{K}$ $\qquad \cdots \quad (7.1.6)$

In the chapter on diodes we have obtained the parameters

$$\left.\begin{aligned} y^*_{11B} &= \frac{1}{re}(1 + j\omega\, T_{11}) \\ y^*_{21B} &= \frac{1}{re}(-\alpha_0 + j\omega\, T_{21}) \end{aligned}\right\} \quad \cdots \quad (7.1.7)$$

We need only to find the parameters y^*_{12B} and y^*_{22B}.

7.2 Early effect

To find the parameters y^*_{12B} and y^*_{22B} in the early days of transistors the effect of collector base voltage variation on the carrier concentration at the collector depletion region boundary on the base side was taken into account exactly alike the emitter boundary but because the collector junction is reverse biased, this variation was small. Later it was found that there is a much more pronounced effect than the depletion region potential variation causing injection or extraction at this boundary. That is when this depletion region potential varies the depletion region width varies causing variation of base width.

In the chapter on diode we assumed this boundary as fixed and the excess carrier concentration there essentially zero but because of the variation of base width with collector base voltage the concentration there would vary as shown in Fig. (7.2.1) by δp. The collector base voltage being time varying dp also becomes time varying. This is known as Early effect.

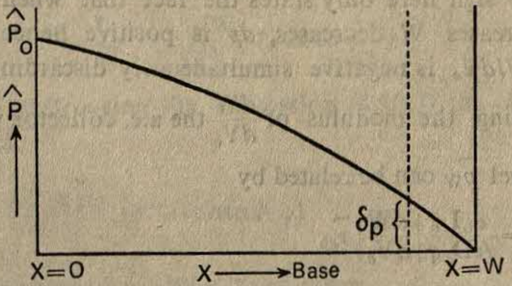


Fig. 7.2.1. Increase of collector junction reverse voltage decreases the base region width and an increase of concentration at the collector junction boundary of the base at dotted line at this instant.

This is equivalent to an injection at the collector boundary and keeping the previously found values for $y^*_{11\mathrm{B}}$ and $y^*_{21\mathrm{B}}$ the emitter current or collector currents can be found out for these injected carriers which, in their turn, would give the values for $y^*_{12\mathrm{B}}$ and $y^*_{22\mathrm{B}}$ respectively.

When we proceed to find these we need to find out the value of dp due to variation of collector voltage. To find this injection level at the collector base depletion boundary due to Early effect, we start from the equation (4.4.3) which reads

$$\mathrm{J}_p = -\mathrm{D}_p q\left(\frac{d\hat{p}}{dx} - \frac{2\theta}{\mathrm{W}}\hat{p}\right) \qquad \cdots \quad (4.4.3)$$

Since at W the field current is zero (being $\hat{p} \simeq 0$) the hole current is

$$\mathrm{J}_{pc} = -\mathrm{D}_p q\frac{d\hat{p}}{d\mathrm{W}} \qquad \cdots \quad (7.2.1)$$

remembering that $\hat{p}$ at W is small we have then

$$dp = \frac{-\mathrm{J}_{pc}}{\mathrm{D}_p q}\cdot\frac{d\mathrm{W}}{d\mathrm{V}_c}\cdot d\mathrm{V}_c = \frac{a_0\mathrm{J}_{pe}}{\mathrm{D}_p q}\left(-\frac{d\mathrm{W}}{d\mathrm{V}_c}\right)\overleftharpoondown{v}_c \qquad \cdots \quad (7.2.2)$$

The negative sign here only states the fact that when the collector voltage increases W decreases, dp is positive hence an injection. Because $d\mathrm{W}/d\mathrm{V}_c$ is negative simultaneously discarding the negative sign and taking the modulus of $\frac{d\mathrm{W}}{d\mathrm{V}_e}$ the a.c. collector voltage $\overleftharpoondown{v}_c$ and injection level $\overleftharpoondown{p}_\mathrm{W}$ can be related by

$$\overleftharpoondown{p}_\mathrm{W} = \frac{a_0\mathrm{I}_{\mu e}}{\mathrm{AD}_p q}\left|\frac{d\mathrm{W}}{d\mathrm{V}_c}\right|\overleftharpoondown{v}_{cb} \qquad \cdots \quad (7.2.3)$$

Either in a *p-n-p* or an *n-p-n* transistor emitter base voltage magnitude increase causes the emitter injection current to increase, this flowing through a resistance load causes collector to base voltage magnitude decrease with consequent increase in base width, resulting in decrease of collector current. In this respect, Early effect implies negative feedback at least at low frequencies. This is viewed as $\overleftharpoondown{v}_{cb}$ positive (note agreement with circuit convention of current) causing injection $\overleftharpoondown{p}_\mathrm{W}$. This injected particle flow is opposite to the flow of particles injected by the emitter. The circuit convention of current flowing in by both emitter and collector is a fitting choice in this respect.

Once we have obtained the a.c. injection level $\overset{\leftharpoonup}{p}_{\mathrm{W}}$ as related to the collector base voltage $\overset{\leftharpoonup}{v}_{cb}$ this $\overset{\leftharpoonup}{p}_{\mathrm{W}}$ can be used to find the currents flowing due to Early effect at the emitter and collector ends.

For this we need not solve the continuity equation again. Instead we can use expression (4.10.6) and (4.10.7) for currents at the collector and the emitter end respectively with suitable modification. Firstly the carriers are now moving opposite to the direction of the built in field we replace θ by $-\theta$, secondly instead of $\frac{q}{kT}\hat{p}_0\overset{\leftharpoonup}{v}_e$ we write $\overset{\leftharpoonup}{p}_{\mathrm{W}}$ the present injection level; thirdly, we take account of the reversal of the direction of current and the reversal of emitter and collector ends. With these in mind we can write for the current at the collector end from expression (4.10.6) as

$$i_{pe\mathrm{W}}=\frac{\mathrm{A}\overset{\leftharpoonup}{p}_{\mathrm{W}}\,\mathrm{D}_p q}{\mathrm{W}}.\,(\psi\coth\psi-\theta) \qquad \cdots \quad (7.2.4)$$

substituting for $\overset{\leftharpoonup}{p}_{\mathrm{W}}$ from eqn. (7.2.3) and noting that this current is in the direction of the circuit convention collector current we have

$$y^*_{22\mathrm{B}}=\frac{i_{pe\mathrm{W}}}{\mathrm{V}_{cb}}=\frac{\mathrm{I}_{pe}a_0}{\mathrm{W}}\left|\frac{d\mathrm{W}}{d\mathrm{V}_c}\right|(\psi\coth\psi-\theta) \qquad \cdots \quad (7.2.5)$$

Similarly, we write using the expression (4.10.7) the current at the emitter end as

$$i_{pco}=\frac{\mathrm{AD}_p q\overset{\leftharpoonup}{p}_{\mathrm{W}}}{\mathrm{W}}\,(e^{-\theta}.\psi\,\mathrm{cosech}\,\psi) \qquad \cdots \quad (7.2.6)$$

which with the value of $\overset{\leftharpoonup}{p}_{\mathrm{W}}$ substituted gives

$$i_{pco}=\frac{\mathrm{I}_{pe}}{\mathrm{W}}\cdot a_0\left|\frac{d\mathrm{W}}{d\mathrm{V}_c}\right|(e^{-\theta}\,\psi\,\mathrm{cosech}\,\psi)\,v_{cb} \qquad \cdots \quad (7.2.7)$$

Remembering that this current is coming out of the emitter terminal and opposite to the circuit convention of emitter current we have

$$y^*_{12\mathrm{B}}=-\frac{i_{pco}}{v_{cb}}=-\frac{\mathrm{I}_{pe}a_0}{\mathrm{W}}\left|\frac{d\mathrm{W}}{d\mathrm{V}_c}\right|(e^{-\theta}\,\psi\,\mathrm{cosech}\,\psi) \qquad \cdots \quad (7.2.8)$$

Rewriting equations (7.1.7)

$$y^*_{11B}=\frac{1}{r_e}\left(\frac{\theta+\psi\coth\psi}{\theta+\psi_0\coth\psi_0}\right)\simeq\frac{1}{r_e}(1+j\omega\, T_{11}) \quad \cdots \quad (7.2.9)$$

$$y^*_{21B}=-\frac{qI_{pe}}{kT}\left(\frac{e^{\theta}\psi\ \mathrm{cosech}\ \psi}{\theta+\psi_0\coth\psi_0}\right)\simeq\frac{1}{r_e}(-a_0+j\omega\, T_{21}) \quad \cdots \quad (7.2.10)$$

We have obtained all the four common base intrinsic admittance parameters. From equation (7.1.6), (7.2.8) and (7.2.10), we have

$$K=\frac{y^*_{21B}}{y^*_{12B}}=\frac{q\omega}{kT}\left|\frac{dV_c}{dW}\right| e^{\theta}.\ \frac{\sinh\psi_0}{\psi_0} \quad \cdots \quad (7.2.11)$$

For the value of $\theta=0$, *i.e.*, for uniformly doped base transistor, the power gain is

$$K_1=\frac{qW}{kT}\left|\frac{dV_c}{dW}\right| \quad \cdots \quad (7.2.12)$$

The factor $\left|\frac{dV_c}{dW}\right|$ can be easily found out, from the expression for the depletion region width '*d*' of a step junction when base doping is very light compared to the collector doping, from equation (3.1.6)

$$d=\left[\frac{2\varepsilon\varepsilon_0\, V_t}{q\, N_d}\right]^{1/2} \quad \cdots \quad (3.1.6)$$

Since in this case the depletion region extends mainly into the base we have

$$\frac{dW}{dV_c}=\frac{-d(d)}{dV_c}=-\left[\frac{\varepsilon\varepsilon_0}{2q\, N_d\, V_t}\right]^{1/2} \quad \cdots \quad (7.2.13)$$

where for reverse bias $V_t=V_b+V_{cb}$.

A typical example if $N_d=10^{15}$/cc and $N_a=10^{16}$, $\left|\frac{dV_c}{dW}\right|=10^5$V/cm and for a base width $W=2.5$ micron the power gain at room temperature becomes $\sim 10^3$.

In a drift transistor with the increase of collector voltage the penetration of the depletion region in base is less than that of an uniform base (diffusion) transistor. This is because increase of voltage brings the depletion region boundary into higher and higher doped region hence lesser penetration into base. Therefore, in a drift transistor $\left|\frac{dV_c}{dW}\right|$ is much greater than in a diffusion transistor.

Over and above for $\theta > 0$ both e^{θ} and $\left(\frac{\sinh \varphi_0}{\varphi_0}\right)$ are greater than unity hence the power gain (cf. equation 7.2.11) is very much greater in a drift transistor and this enables present day transistors to have lower base width ~0.5 micron still having power gain 10,000 or so. Less base width, in turn, gives less base transit time and so higher frequency of operation. The highest order of magnitude of K occurs for $\theta = 4$, $W/L_p \leqslant 1$ and if $W = 0.5$ micron, then $|dV_o/dW| = 10$ volts per micron approaching the breakdown field 200 KV/cm. Lastly, approximate expressions can be written for y^*_{12B} and y^*_{22B} from equation (7.2.10) with (7.1.6)

$$y^*_{12B} = \frac{1}{Kr_e}\left(-\alpha_0 + j\omega T_{21}\right) \qquad \cdots \quad (7.2.14)$$

From equations (7.2.4) with (7.2.6) define

$$(i_{pco}/i_{pew}) = \alpha_{0i} = \frac{e^{-\theta}\varphi_0 \operatorname{cosech} \varphi_0}{\varphi_0 \coth \varphi_0 - \theta} \qquad \cdots \quad (7.2.15)$$

It is to be remembered that the reverse α_{0i} is appreciably lower than the forward α_0 because of lower emitter area than the collector electrode also.

Rewrite equation (7.2.5) with (7.2.11) and (7.2.15) as

$$y^*_{22B} = \frac{\alpha_0}{r_e K \alpha_{0i}}\left(1 + j\omega T_{22}\right) \qquad \cdots \quad (7.2.16)$$

where

$$T_{22} = \frac{\alpha_{0i}}{\alpha_0} \cdot e^{2\theta} . T_{11} \qquad \cdots \quad (7.2.17)$$

7.3 Intrinsic α cut-off frequency

By intrinsic α cut-off frequency, we mean the α cut-off frequency of the intrinsic transistor. We have also mentioned how this ultimately turns out to be the base region transport factor (cf. art. 6.3). This is again from equation (4.11.1)

$$\alpha = \frac{\varphi e^{\theta}}{\theta \sin \varphi + \varphi \cosh \varphi} \qquad \cdots \quad (7.3.1)$$

For a certain transistor a value of θ may be taken and a value of $\frac{W}{L_p} \leqslant 1$ can be taken and the value of $\varphi_0 = (\theta^2 + W^2/L_p^2)$ can be computed. Since $\psi^2 = \left(\psi_0{}^2 + j\omega \frac{W^2}{D_p}\right)$, a value for $\omega \frac{W^2}{D_p}$ can be assumed and the value of ψ can be found as a complex quantity $(a + jb)$. The cosh ψ and sinh ψ terms can be expanded as $\cosh \psi = \cosh a \cos b + j \sinh a \sin b$ and $\sinh \psi = \sinh a \cos b + j \cosh a \sin b$

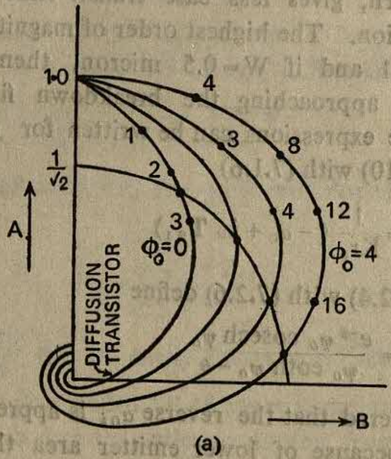
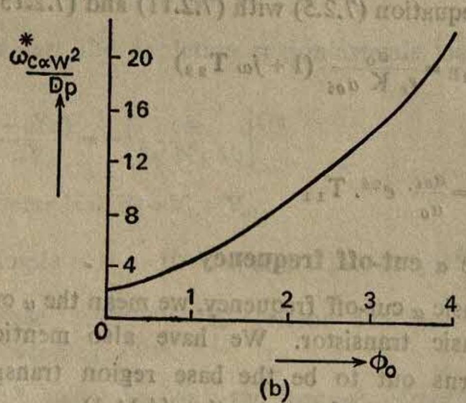


Fig. 7.3.1. (a) Complex plot of (α/α_0)



(b) ω W²/D$_p$ plot versus ψ_0

and the value of α can be computed. The value of α_0 can also be computed easily. The ratio (α/α_0) can be found out in the form A + jB for the assumed $\omega \frac{W^2}{D_p}$. A complex plane plot of the this can be made with the value of $\omega \frac{W^2}{D_p}$, marked on the point as shown in Fig. (7.3.1*a*). Similar curves can be obtained for different values of ψ_0 ranging from very small to the value 4. The intersection of each curve with a circle of radius 0.707 gives the value of $\omega \frac{W^2}{D_p}$ which gives the corresponding cut-off frequency for this value ψ_0. If ω^*_{ca} is the cut-off frequency then $\omega^*_{ca} \frac{W^2}{D_p}$ corresponding to each value of ψ_0 can be read off from the marks on the curves and plotted against ψ_0 as shown in Fig. (7.3.1*b*). Joining these points and by trial and error one can find as equation for this curve as

$$\omega^*_{ca} \frac{W^2}{D_p} = \frac{\pi^2}{4} (1 + \psi_0{}^{4/3}) \qquad \cdots \quad (7.3.2)$$

which in turn gives

$$\omega^*_{ca} = \frac{\pi^2}{4} \frac{D_p}{W^2} (1 + \psi_0{}^{4/3}) \qquad \cdots \quad (7.3.3)$$

for the intrinsic α cut-off frequency as related to the built-in field, since, $\theta \simeq \psi_0$ and $E = \frac{2\theta}{W} \frac{kT}{q}$.

From the above consideration, a first approximation to the equation (7.3.1) can be written as

$$\alpha = \frac{\alpha_0}{\left[1 + i \frac{\omega}{\omega^*_{ca}}\right]} \qquad \cdots \quad (7.3.4)$$

The phase from this expression at ω^*_{ca} is 45° lagging. While the diagram shows that the phase increases with ψ_0. For each value of ψ_0 at ω^*_{ca} the phase can be read from Fig. (7.3.1*a*) and the value less 45° can be plotted as a function of ψ_0 to give the extra bit of phase lag that can be incorporated in (7.3.4) ; hence a little more computation indicates, α can be expressed as

$$\alpha = \left[\alpha_0 \Big/ \left(1 + j \frac{\omega}{\omega^*_{ca}}\right)\right] \exp\left[-j \frac{\pi}{15} (1 + \psi_0) \frac{\omega}{{}^*\omega_{ca}}\right] \qquad \cdots \quad (7.3.5)$$

7.4 Extrinsic cut-off frequency

In this section we shall consider how the external parameters such as C_e, C_c and r_b see Fig. (7.4.1), would modify this intrinsic cut-off frequency and how the overall cut-off frequency is related to these external parameters and the intrinsic ones.

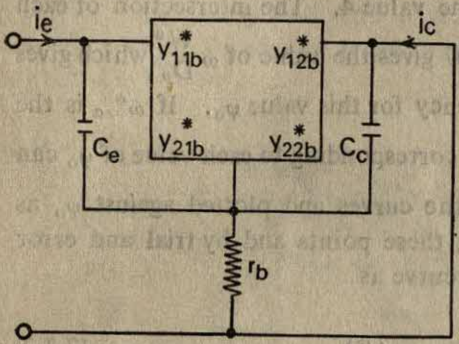


Fig. 7.4.1. Showing the output shorted transistor in C.B.

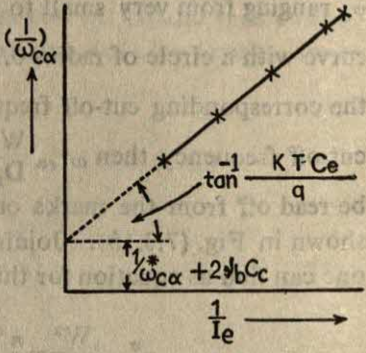


Fig. 7.4.2. Shows the plot of $(1/I_e)$ versus $(1/\omega_{c\alpha})$

When we think of the overall transistor cut-off frequency, we need to find out $\left(\frac{i_c}{i_e}\right)$ when the base collector has an a.c. short. Using the common base intrinsic parameters one can easily show for the overall transistor

$$\alpha=\frac{-y^*_{21b}+[r_b\,\{\Delta y^*_b+j\omega\,C_c\,(y^*_{11b}+j\omega C_c)\}+j\omega C_c\,y^*_{22b}]}{y^*_{11b}+j\omega C_e+[r_b\,\{\Delta y^*_b+j\omega C_c\,(y^*_{11b}+j\omega C_c)\}+j\omega C_c\,y^*_{22b}]} \quad \cdots \quad (7.4.1)$$

where $\Delta y^*_b=(y^*_{11b}\,y^*_{22b}-y^*_{12b}\,y^*_{21b})$ $\cdots$ (7.4.2)

Since it is practically impossible from the expression (7.4.1) to visualise how the different extrinsic parameters are controlling the overall α, we resort to a rough approximation as a guide to have an analytical result. In doing this, we ignore terms containing y^*_{22b} and y^*_{12b} since they are small by there having K, a large quantity in the denominator and approximating $\frac{j\omega C_e}{y^*_{11b}} \simeq j\omega\, r_e C_e$, we have on simplification

$$\alpha=\frac{\left(\frac{\alpha^*}{1+j\omega\,C_e\,r_e}+j\omega r_b\,C_c\right)}{1+j\omega\,r_b\,C_c}. \quad \cdots \quad (7.4.3)$$

Considering $\alpha_0=1$, ignoring the phase factor $\left[\frac{\pi}{15}(1+\varphi_0)\frac{\omega}{\omega^*_{c\alpha}}\right]$ and assuming $r_b C_c << \left(C_e r_e + \frac{1}{\omega^*_{ca}} + r_b C_c\right)$, it can be easily shown that the α cut-off radian frequency ω_{ca} can be written as

$$\frac{1}{\omega_{ca}} = \frac{1}{\omega^*_{ca}} + r_e C_e + r_b C_c \qquad \cdots \quad (7.4.4)$$

Taking recourse to computational method, Lyndmayer (cf. ref. 1) finds from (equation 7.4.1)

$$\frac{1}{\omega_{ca}} = \frac{1}{\omega^*_{ca}} + r_e C_e + 2 r_b C_c \qquad \cdots \quad (7.4.5)$$

To get a feel of the extrinsic parameters modifying effect on the intrinsic alpha cut-off frequency, let us consider a typical problem. A transistor has $\theta = 2$, assume $D_p = 50$ cm^2/sec and $W = 1$ micron. Since $\frac{W}{L_p} << 1$ taking $\theta = \psi_0 = 2$, we have $f^*_{ca} = 6.9$ GHz. Assuming $C_e = 15$ *pf* $r_b = 100$ Ohms, $C_c = 1$ *pf*, $r_b C_e$ time constant reduces the f_{ca} to 715 Mc/s while $r_e C_e$ gives a strong current dependence. At 1 mA, the overall $f_{ca} = 266$ MHz, at 6 mA 556 MHz. Thus we see that in a practical transistor the external parameters reduce the overall cut-off frequency by factors of 10 to 20. At these frequencies the intrinsic transistor is basically working in its low frequency mode.

Equation (7.4.5) can be written as

$$\frac{1}{\omega_{ca}} = \left[C_e \frac{kT}{qI_e} + \left(2r_b C_c + \frac{1}{\omega^*_{ca}}\right)\right] \qquad \cdots \quad (7.4.6)$$

$(1/\omega_{ca})$ can be measured for different values of $(1/I_e)$ and plotted as shown in Fig. (7.4.2). This is a straight line which can be extrapolated to yield $\frac{kTC_e}{q}$; hence C_e is obtained from the slope and $(1/\omega^*_{ca} + 2r_b C_c)$ as the intercept on $1/\omega_{ca}$ axis at $1/I_e \rightarrow 0$.

7.5 Common base extrinsic parameters

In this section we shall find out the common base extrinsic parameters using the CB intrinsic parameters and the external elements C_e and C_c as shown in the Fig. (7.5.1).

For the network on the right the small signal current voltage relations can be written as

$$\left.\begin{aligned} i_e &= y_{11b}\, v_{be} + y_{12b}\, v_{be} \\ i_c &= y_{21b}\, v_{be} + y_{22b}\, v_{bc} \end{aligned}\right\} \quad \cdots \quad (7.5.1)$$

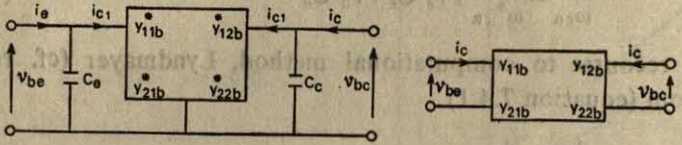


Fig. 7.5.1. Shows the equivalent newtworks for inclusion of extrinsic parameters into the Black Box

whereas for the intrinsic active network on the left, we can write

$$\left.\begin{aligned} i_{e1} &= y^*_{11b}\, v_{be} + y^*_{12b}\, v_{bc} \\ i_{c1} &= y^*_{21b}\, v_{be} + y^*_{22b}\, v_{bc} \end{aligned}\right\} \quad \cdots \quad (7.5.2)$$

From equations (7.5.1), (7.5.2) and fig. (7.5.1) for $v_{bc}=0$; $i_c=i_{c1}$

$$y_{11b}=\frac{i_e}{v_{be}}=\frac{i_{e1}+i_1}{v_{be}}=\frac{y^*_{11b}\, v_{be}+j\omega C_e\, v_{be}}{v_{be}}=y^*_{11b}+j\omega C_e \quad \cdots \quad (7.5.3)$$

$$y_{21b}=\frac{i_c}{v_{be}}=\frac{i_{c1}}{v_{be}}=y^*_{21b} \quad \cdots \quad (7.5.4)$$

Similarly, when $v_{be}=0$, $i_e=i_{e1}$ and from eqns. (7.5.1) and (7.5.2) we have

$$y_{22b}=\frac{i_c}{v_{bc}}=\frac{i_{c1}+i_2}{v_{bc}}=y^*_{22b}+j\omega C_c \quad \cdots \quad (7.5.5)$$

$$y_{12b}=\frac{i_e}{v_{bc}}=\frac{i_{e1}}{v_{bc}}=y^*_{12b} \quad \cdots \quad (7.5.6)$$

7.6 Common Emitter Parameters

In this article we shall find out the common emitter extrinsic parameters from the common base extrinsic parameters of the previous article. Since in the common emitter configuration the base is the input terminal, the base resistance can be included in series with the base terminal later on.

Fig. (7.6.1) shows the common base and the common emitter configuration y parameters with input and output currents and

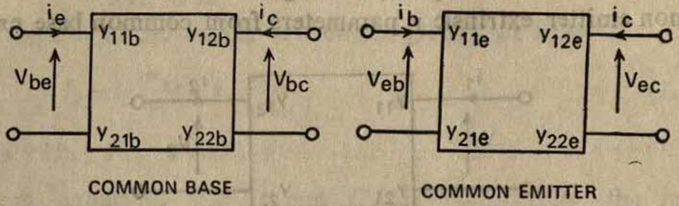


Fig. 7.6.1 Common Base and Common Emitter

voltages marked. Since in circuit convention all the currents are entering we have $i_e+i_b+i_c=0$ and the voltage added vectorially

$$v_{be}+v_{ec}-(v_{be})=0 \text{ i.e., } v_{be}=-v_{eb} \text{ etc.} \qquad \cdots \quad (7.6.1)$$

The common emitter eqn. can be written as

$$\left.\begin{aligned} i_b &= y_{11e}\,v_{eb}+y_{12e}\,v_{ec} \\ i_c &= y_{21e}\,v_{eb}+y_{22e}\,v_{ec} \end{aligned}\right\} \qquad \cdots \quad (7.6.2)$$

While the common base equns. can be written as

$$\left.\begin{aligned} i_e &= y_{11b}\,v_{oe}+y_{12b}\,v_{bc} \\ i_c &= y_{21b}\,v_{be}+y_{22b}\,v_{bc} \end{aligned}\right\} \qquad \cdots \quad (7.6.3)$$

From (7.6.3) with (7.6.2) we have

$$\begin{aligned} i_b = -(i_e+i_c) &= (y_{11b}+y_{21b})\,v_{eb}-(y_{12b}+y_{22b})\,(v_{be}+v_{ec}) \\ &= (y_{11b}+y_{12b}+y_{21b}+y_{22b})\,v_{eb}-(y_{12b}+y_{22b})\,v_{ec} \end{aligned} \qquad \cdots \quad (7.6.4)$$

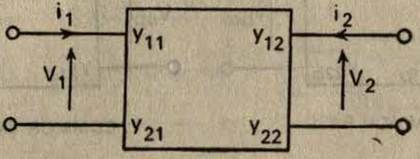
Again from (7.6.3) with (7.6.2), we have,

$$\begin{aligned} i_c &= y_{21b}\,v_{be}+y_{22b}\,v_{bc} = -y_{21b}\,v_{eb}-y_{22b}\,(v_{bc}+v_{ec}) \\ &= -(y_{21b}+y_{22b})\,v_{eb}+y_{22b}\,v_{ec} \end{aligned} \qquad \cdots \quad (7.6.5)$$

Comparing appropriate eqns. (7.6.2) with (7.6.4) and of (7.6.2) with (7.6.5) we can write

$$\left.\begin{aligned} y_{11e} &= y_{11b}+y_{12b}+y_{21b}+y_{22b} \\ y_{12e} &= -(y_{12b}+y_{22b}) \\ y_{21e} &= -(y_{21b}+y_{22b}) \\ y_{22e} &= y_{22b} \end{aligned}\right\} \qquad \cdots \quad (7.6.6)$$

Equations (7.5.3) to (7.5.6) give the formulae to be used for obtaining the common base extrinsic y parameters from common base intrinsic y parameters. While eqn. (7.6.6) gives the formulae to obtain common emitter extrinsic y parameters from common base extrinsic y parameters. For quick reference, common base intrinsic y parameters can be written from eqns. (7.2.9, 7.2.10 and 7.2.14) as



Fig, 7.7.1. A general two-part network

$$\left.\begin{aligned} y^*_{11b} &= \frac{1}{r_e}(1+j\omega\,T_{11}) \\ y^*_{21b} &= \frac{1}{r_e}(-\alpha_0+j\omega\,T_{21}) \\ y^*_{12b} &= \frac{1}{Kr_e}(-\alpha_0+j\omega\,T_{21}) \\ y^*_{22b} &= \frac{1}{Kr_e}\,\frac{\alpha_0}{\alpha_{0i}}(1+j\omega\,T_{22}) \end{aligned}\right\} \qquad \cdots \quad (7.6.7)$$

and from equations (7.5.3) to (7.5.6)

$$\left.\begin{aligned} y_{11b} &= y^*_{11b}+j\omega\,C_e \\ y_{21b} &= y^*_{21b} \\ y_{12b} &= y^*_{12b} \\ y_{22b} &= y^*_{22b}+j\omega\,C_e \end{aligned}\right\} \qquad \cdots \quad (7.6.8)$$

7.7 Equivalent Circuit

In this section we discuss how we can obtain the equivalent circuit of the transistor. This section does not confine to any specific transistor configuration and so refers to the physical quantities in terms of y parameters of the particular configuration and so the subscripts b or e have been dropped.

From the figure (7.7.1), we can write the current equations as

$$\left.\begin{aligned} i_1 &= y_{11}\,v_1+y_{12}\,v_2 \\ i_2 &= y_{21}\,v_1+y_{22}\,v_2 \end{aligned}\right\} \qquad \cdots \quad (7.7.1)$$

The first of these equations can be written as

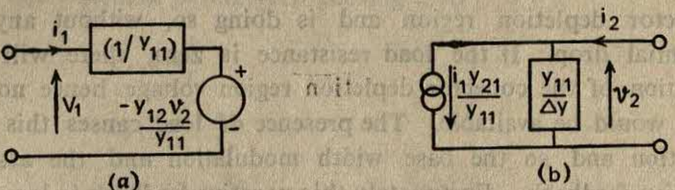
$$v_1 = \frac{i_1}{y_{11}} - \frac{y_{12}}{y_{11}} v_2 \qquad \cdots \quad (7.7.2)$$

Substituting this in the second equation of (7.7.1), we have

$$i_2 = i_1 \frac{y_{21}}{y_{11}} + v_2 \left(\frac{\Delta y}{y_{11}}\right) \qquad \cdots \quad (7.7.3)$$

where $\Delta y = (y_{11}\, y_{22} - y_{12}\, y_{21})$ $\qquad \cdots \quad (7.7.4)$

From equations (7.7.2) and (7.7.3) respectively the input equivalent circuit and the output equivalent circuit can be set up as in Fig. (7.7.2).



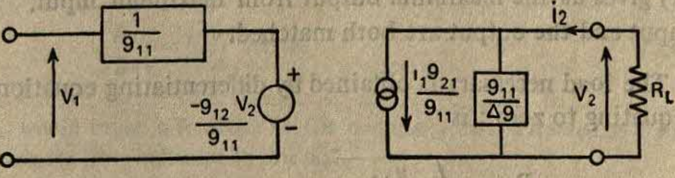
i_1
$(1/y_{11})$
V_1
+
$-\frac{y_{12}v_2}{y_{11}}$
−
(a)
i_2
$\frac{i_1 y_{21}}{y_{11}}$
$\frac{y_{11}}{\Delta y}$
v_2
(b)

Fig. 7.7.2. Input and output equivalent circuits applicable irrespective of configuration

In Fig. (7.7.2a), the voltage v_1 is the sum of two voltages, one due to drop in impedance $(1/y_{11})$ and the other, a feedback factor $(-y_{12}/y_{11})\, v_2$. In Fig. (7.7.2*b*), the current is the sum of two currents, one due to a current generator and that in the impedance $(y_{11}/\Delta y)$ due to v_2.

To set up the equivalent circuit in any particular transistor configuration, use will be made of equations (7.6.7) and (7.6.8) for common base configuration and (7.6.6) with (7.6.8), and (7.6.7) for common emitter configuration.

At very low frequencies, all the y parameters can be replaced by their conductance parts because the reactance parts tend to zero



$\frac{1}{g_{11}}$
V_1
+
$\frac{-g_{12}V_2}{g_{11}}$
−
i_2
$\frac{i_1 g_{21}}{g_{11}}$
$\frac{g_{11}}{\Delta g}$
V_2
R_L

Fig. 7.7.3 Input and output equivalent circuits for low frequency

and in that case the equivalent circuit can be formed as shown in fig. (7.7.3) where g_{11} is the conductance part of y_{11} and so on; while

$$\Delta g=(g_{11}\,g_{22}-g_{12}\,g_{21}) \qquad \cdots \quad (7.7.5)$$

7.8 Low frequency power gain

The low frequency power gain of the transistor was found as K that comes from the voltage, varying across the collector depletion region. The current is controlled by emitter injection and recombination in base. The base collector junction, being reverse biased, has no control over this current though it is flowing through the collector depletion region and is doing so, without any extra potential drop. If the load resistance is zero, there will be no variation of the collector depletion region voltage hence no power gain would be available. The presence of load causes this voltage variation and so the base width modulation and the associated negative feedback. Fortunately this negative feedback is low.

In this section we examine what is the maximum available power gain of the transistor. For a certain load R_L the transistor power gain is obtained as

$$\text{Power gain}=\frac{P_o}{P_{in}}=\frac{R_L\,g_{21}{}^2}{(g_{11}+R_L\,\Delta g)(1+R_L\,g_{22})} \qquad \cdots \quad (7.8.1)$$

This power gain has been calculated in terms of R_L. This is for a certain R_L, the output voltage hence the feedback voltage has certain value in terms of input current and hence the input power is obtained, while the output power is determined by the input current and R_L. This power gain would vary if R_L is varied because of the reasons : the input alone can get matched or the output alone can get matched or both input and output can get matched simultaneously. Maximisation of power gain given by equation (7.8.1) gives us the maximum output from maximum input. That is the input and the output are both matched.

The load necessary is obtained by differentiating equation (7.8.1) and equating to zero ; is

$$R_L=\sqrt{\frac{g_{11}}{\Delta g.\,g_{22}}} \qquad \cdots \quad (7.8.2)$$

Under this matched condition the generator resistance is

$$R_g = \sqrt{\frac{g_{22}}{\Delta g \cdot g_{11}}} \quad \cdots \quad (7.8.3)$$

and is equal to the input resistance.

With input and output matching, the maximum available power gain is obtained as

$$G_{\max} = \frac{{g_{21}}^2}{\Delta g \left[1 + \sqrt{\frac{g_{11}\, g_{22}}{\Delta g}}\right]^2} \quad \cdots \quad (7.8.4)$$

From eqns. (7.6.8) and (7.6.7), we can write the low frequency common base extrinsic parameters ignoring r_b as*

$$\left.\begin{aligned} g_{11b} &= \frac{1}{r_e} \\ g_{21b} &= -\frac{a_0}{r_e} \\ g_{12b} &= -\frac{a_0}{\mathrm{K} r_e} \\ g_{22b} &= \frac{1}{\mathrm{K} r_e}\,\frac{a_0}{a_{0i}} \\ \text{and} \qquad \Delta g &= \frac{a_0(1 - a_0\, a_{0i})}{a_{0i}\, \mathrm{K} {r_e}^2} \end{aligned}\right\} \quad \cdots \quad (7.8.5)$$

The maximum available *l.f.* common base power gain, as obtained from eqn. (7.8.4) with eqn. (7.8.5), is

$$(G_{\max})_{CB} = \frac{a_0\, a_{0i}\, \mathrm{K}}{(1 - a_0\, a_{0i})\left[1 + \sqrt{\frac{1}{1 - a_0\, a_{0i}}}\right]^2} \simeq \mathrm{K} \quad \cdots \quad (7.8.6)$$

since both a_0 and a_{0i} have near unity values.

The ratio of output to input impedance under matched condition is

$$\frac{g_{11}}{g_{22}} = \mathrm{K}\,\frac{a_{0i}}{a_0} \simeq \mathrm{K} \quad \cdots \quad (7.8.7)$$

N.P. r_b would cause a feedback in CB and easily can be treated as a circuit problem along the lines indicated in this text.

Therefore for cascading CB stages, step down transformer having turns ratio of $\sqrt{K}$ has to be used for matching input of second stages to the output of first stage for maximum power transfer. Consequently, if the input and output are matched the voltage gain per stage is only $\sqrt{K}$.

7.9 Unilateral power gain

Unilateral power gain is the power gain when power flows in the forward direction only. That is in the present case if the feedback generator is eliminated unilateralisation occurs. Assuming that the feedback parameter g_{12} is absent *i.e.*, equal to zero the unilateral power gain can be calculated. For output matching (cf. Fig. 7.7.3), $R_L = g_{11}/\Delta g$ so half of the current of the current generator flows through this load to give the output power

$$P_{out} = \left(\frac{g_{21}}{g_{11}} \; \frac{i_1}{2}\right)^2 \cdot \; \frac{g_{11}}{\Delta g} \qquad \cdots \quad (7.9.1)$$

while the input power is $P_{in} = \dfrac{i_1^2}{g_{11}}$ $\qquad \cdots \quad (7.9.2)$

Hence power gain $U = \dfrac{g_{21}^2}{4\Delta g}$ $\qquad \cdots \quad (7.9.3)$

From (7.9.3) and (7.8.5), we have, for the intrinsic CB transistor, unilateral power gain

$$U = \frac{K a_0 \, a_{0i}}{4(1 - a_0 \, a_{0i})} \qquad \cdots \quad (7.9.4)$$

From the common emitter configuration from equations (7.6.6) with (7.8.5), ignoring terms containing frequency, we have

$$\left.\begin{aligned} g_{11e} &= \frac{1 - a_0}{r_e} \\ g_{12e} &= -\frac{a_0}{a_{0i}\, K r_e}\,(1 - a_{0i}) \\ g_{21e} &= \frac{a_0}{r_e} \\ g_{22e} &= \frac{a_0}{a_{0i}\, K r_e} \\ \text{and} \quad & \\ \Delta g_e &= a_0\,(1 - a_0\, a_{0i})/(a_{0i} \cdot K r_e^2) \end{aligned}\right\} \qquad \cdots \quad (7.9.5)$$

From eqn. (7.9.3) with (7.9.5), we see that the unilateralised CE power gain is

$$\frac{K}{4}\ \frac{\alpha_0\,\alpha_{0i}}{(1-\alpha_0\,\alpha_{0i})} \qquad \cdots \quad (7.9.6)$$

Thus we see that the unilateral power gain *i.e.*, the maximum gain available of a transistor having no internal feedback is same (cf. equation 7.9.4 and 7.9.6) in both the configuration. This is true so long g_{12} hence the feedback is ignored. In this respect the common emitter $g_{12e}=-\frac{\alpha_0}{\alpha_{0i}\,Kr_e}(1-\alpha_{0i})$ is much less than common base $g_{12b}=-\frac{\alpha_0}{Kr_e}$ and so unilateralisation is nearly there in common emitter configuration. From the above consideration, it can be said that in CE configuration the power gain is higher than in CB configuration because in CE, more near unilateralisation exists.

At this stage it is worth considering input and output impedance levels. Since g_{12}, in either configuration, is low because primarily of high value of K, it is quite in order if g_{12} is ignored. The input impedance of the two configuration CE & CB can be compared ignoring r_b. The input resistance under the above restriction is $1/g_{11}$ and the output resistance $g_{11}/\Delta g$. This gives (c.f. equations 7.8.5 and 7.9.5) for CB configuration, the input resistance $r_e=(1/g_{11b})$ small while for CE configuration, input resistance $\frac{r_e}{1-\alpha_0}\simeq\beta_0\,r_e$.

$$\left.\begin{array}{l}\text{As for output resistances in CB, } \frac{Kr_e}{\alpha_0}\left[\frac{\alpha_{0i}}{1-\alpha_0\,\alpha_{0i}}\right]\\ \textit{i.e.}\text{, very large ; while for CE configuration it is}\\ \quad \frac{Kr_e}{\beta_0}\left(\frac{\alpha_{0i}}{1-\alpha_0\,\alpha_{0i}}\right)\text{, lower than in CB}\end{array}\right\} \cdots \quad (7.9.7)$$

Thus we see that in the common base configuration the input resistance is low and the output resistance is very high whereas in the common emitter configuration, the input resistance is of a moderate value (for 1mA and $\beta_0=100$, *i.e.*, $\alpha_0=0.99$, $\beta_0 r_e=2.5$ K-Ohms) while the output impedance is also moderate for $K=10^4$ and $r_e=25$ Ohms, $\beta_0=100$ and $\beta_{0i}=20$ it is 50K Ohms, whereas the common base value 5×10^6 Ohms. The CE configuration has more moderate impedance levels, so this is preferred by circuit engineers.

7.10 High frequency consideration in CE configuration

In this section the CE high frequency parameters and equivalent circuit will be considered. From eqns. (7.6.6), (7.6.7) and (7.6.8) we can write the high frequency CE extrinsic parameters neglecting terms containing K in the denominator as

$$\left.\begin{aligned} y_{11e} &= \frac{1}{r_e}\,[(1-a_0)+j\omega\,(T_b+C_e\,r_e+C_c\,r_e)] \\ y_{12e} &= -\,j\omega C_c \\ y_{21e} &= -\frac{1}{r_e}\,(-\,a_0+j\omega T_{21})-j\omega C_c \\ y_{22e} &= j\omega C_c \end{aligned}\right\} \quad \cdots \quad (7.10.1)$$

From (7.10.1) we can easily find out the high frequency equivalent circuit parameters and can frame the equivalent circuit according to Fig. (7.7.2). The base resistance in CE consideration is not common to input and output ; this can be put in series at the input.

Using eqn. (7.10.1) implies that the value of K has been assumed very high *i.e.*, the output impedance very high. Considering that for wide band design, gain per stage can be kept small, the value of load resistance will be small ~ 1K, 2K. Ohms ; hence this approximation should be proper. This approximation shall not give an accurate value for the output impedance but would be adequate for considering the roll of device parameters in influencing high frequency performance.

7.11 β Cut-off frequency

β is the common emitter short circuit current gain and this can be written as

$$\beta=\frac{y_{21e}}{y_{11e}} \quad \cdots \quad (7.11.1)$$

Substituting values from eqn. (7.10.1)

$$\beta=\frac{y_{21e}}{y_{11e}}=\frac{a_0-j\omega T_{21}-j\omega C_c\,r_e}{(1-a_0)+j\omega\,[T_b+r_e\,(C_e+C_c)]} \quad \cdots \quad (7.11.2)$$

remembering $T_b=T_{11}+T_{21}$; β can be written as

$$\beta=\frac{a_0-j\omega\,(T_b-T_{11}+r_e\,C_e)}{(1-a_0)+j\omega\,[T_b+r_e\,(C_e+C_c)]} \quad \cdots \quad (7.11.3)$$

Equation (7.11.3) can be written, since $\alpha_0 \simeq 1$, as

$$\beta = \frac{\alpha_0}{1-\alpha_0} \frac{1 - j\omega\,(T_b - T_{11} + r_e\,C_c)}{1 + j\omega\,(\beta_0 + 1)\,[T_b + r_c\,(C_e + C_c)]} \qquad \cdots \quad (7.11.4)$$

Considering that the time constant in the denominator is very much larger than that in the numerator of (7.11.4), the frequency variation of β shall be influenced by the denominator time constant much earlier in frequency scale than due to the numerator time constant. Additionally, the numerator time constant can be ignored because when we speak of cut-off frequency, squares of these terms would be compared to find their effect on β magnitude. So, we write eqn. (7.11.4) as

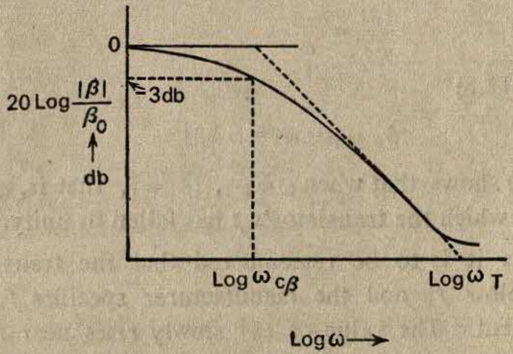


Fig. 7.11.1 Variation of β with frequency (Note Log-Log scale)

$$\beta = \frac{\beta_0}{1 + j\omega/\omega_{c\beta}} \qquad \cdots \quad (7.11.5)$$

where

$$\frac{1}{\omega_{c\beta}} = \beta_0\,[T_b + r_e\,(C_e + C_c)] \qquad \cdots \quad (7.11.6)$$

The time constant $[T_b + r_e\,(C_e + C_c)]$ is dependent on emitter current otherwise it is a constant of the device since it is made up of the transit time through the base, C_e and C_c. For a certain operating current of the transistor, it is a device constant and is denoted by $(1/\omega_T)$. That is, we have in addition,

$$\frac{1}{\omega_T} = \left[T_b + \frac{kT}{q\,I_e}\,(C_e + C_c)\right] \qquad \cdots \quad (7.11.7)$$

From eqn. (7.11.6) and (7.11.7), it is seen that

$$\omega_T = \beta_0 \, \omega_{c\beta} \qquad \cdots \quad (7.11.8)$$

Eqn. (7.11.5) is plotted in Fig. (7.11.1), where it is seen that at $\omega = \omega_{c\beta}$, the value of β falls by 3 *d*B from its low frequency value and at very high frequency, β falls by 6*d*B per octave. In this frequency range, $\omega/\omega_{c\beta} \gg 1$ and eqn. (7.11.5) can be written as

$$\frac{|\beta|}{\beta_0} = \frac{\omega_{c\beta}}{\omega}$$

with eqn. (7.11.8)

$$\beta_0 \, \omega_{c\beta} = \omega_T = \omega \, |\beta| \qquad \cdots \quad (7.11.9)$$

Eqn. (7.11.9) shows that when $\omega = \omega_T$, $|\beta| = 1$, that is, ω_T is a radian frequency at which the transistor's β has fallen to unity.

In this it is to be remembered that the transistor is used normally below f_T and the manufacturer specifies f_T for its ease of measurement. The value of $|\beta|$ slowly rises near f_T because of the time constant in the numerator that has been ignored of eqn. (7.11.4). At this high frequency, the collector depletion region transit time of carriers also becomes important which has been neglected in the operative range of the transistor.

7.12 Measurement of transistor parameters

In this section, the measurement of the transistor small signal parameters f_T, T_b, C_e, C_c & r_b will be considered. For the measurement of f_T & T_b, one needs only to measure $|\beta|$ as a function of frequency. A typical circuit shall be of the form as shown in the Fig. (7.12.1). D.C. current and *d.c.* voltage can be set by potentiometers P_1 & P_2 ; R_1 & R_2 are acting as potential divider giving a low impedance small voltage source for driving the transistor base current, which can be computed from the value of R_3 & R_2, the transistor input impedance being neglected in comparison to R_3. Measurement of voltage v_2 by a V.T.V.M. gives the a.c. voltage developed across the 30 Ohm load which is a practical short circuit. Thus the collector current is obtained and hence the ratio of collector current to base current can be computed knowing the voltage v_1 at the output of the signal generator. C_1 and C_2 are bypass condensers.

If $v_1 = 5$ volts with the typical values quoted in the inset, voltage across R_2 is 50 mv which gives base current of the order of 1 μA, the a.c. base emitter voltage at 0·5 mA d.c. emitter current with $\beta_0 = 100$ is about 5 mv, which ensures that this is much smaller than kT/q.

Thus for a certain I_E & V_{CE} the value of β can be experimentally determined up to values of frequency $\omega \geqslant \omega_{c\beta}$ and plotted as shown

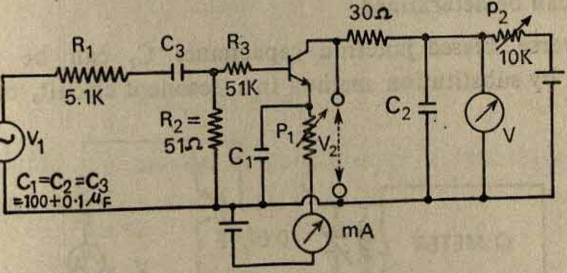


Fig. 7.12.1. Circuit for measuring β vs. frequency

in fig. (7.11.1). The straight line portion is extended to the point $(-20 \log_{10} \beta_0)$ dB and the corresponding frequency is ω_T.

Thus for different values of I_e and V_{CE}, ω_T can be experimentally determined. To determine T_b or $(C_e + C_c)$ as a function of V_{CE}, for each V_{CE} one can plot $1/I_e$ vs. $1/\omega_T$ as shown in fiig. (7.12.2) [cf. eqn. (7.11.7)], extrapolate the straight line portion to obtain $1/\omega_T = T_b$ at $1/I_e - 0$ and the slope gives $(kT/q)\,(C_e + C_c)$. Thus for different V_{CE}, determination of $(C_e + C_c)$ enables one to find the effect

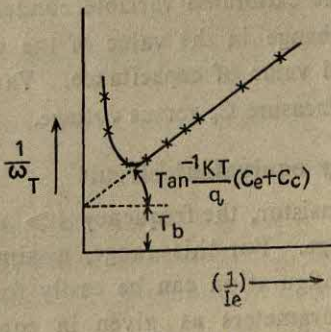


Fig. 7.12.2. Plot of $\frac{1}{\omega_r}$ vs. $\left(\frac{1}{I_e}\right)$ to obtain T_b etc.

of V_{CE} on $(C_e + C_c)$. For a constant emitter current with increase of collector voltage, the collector depletion region width increases, the base width narrows ; therefore, T_b decreases as also C_c ; hence ω_T increases with V_{CE}.

Measurement of base to emitter voltage in fig. (7.12.1) by a V.T.V.M. at very high frequency such that $\omega \gg \frac{1}{r_e C_e}$, shall give the voltage drop in r_b. The current through r_b in Fig. (7.12.1) is known hence r_b can be determined.

Reverse biased junction capacitance C_c can be successfully measured by substitution method in a resonant circuit, or a Q meter

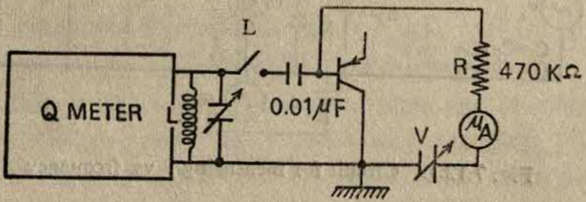


Fig. 7.12.3. Circuit for measurement of reverse biased junction capacitance

as shown in Fig. (7.12.3). Resistance R is very large compared to the reactance of the capacitor being measured. The microammeter gives current hence computation of reverse bias voltage can be done from R and d.c. supply voltage V. If the d.c. supply is derived from a battery eliminator, then the earth of the battery eliminator and that of the Q-meter should be one and the same. The earth remains connected, the resonant system is tuned, the live terminal L is then connected and the calibrated variable condenser is readjusted for resonance. The change in the value of the variable condenser C gives the measured value of capacitance. Variable d.c. voltage supply enables one to measure C_c versus voltage.

7.13 High frequency equivalent circuit

For a certain transistor, the frequency $\omega \gg \omega_{e\beta}$, is regarded as its high frequency range. For this range, a simplified equivalent circuit in the C.E. configuration can be easily found by using the common emitter y parameters as given in eqn. (7.10.1) in the equivalent circuits of fig. (7.7.2).

Substituting $(1/\omega_T)$ for $[T_b + r_e\ (C_e + C_e)]$ in the expression for y_{11} of eqn. (7.10.1), we have

$$y_{11e} = \frac{1}{r_e} \cdot \left[(1-\alpha_0) + j\, \frac{\omega}{\omega_T} \right] \qquad \cdots \quad (7.13.1)$$

from which

$$\frac{1}{y_{11e}} = \frac{(r_e/\overline{1-\alpha_0})}{\left[1 + j\frac{\omega}{\omega_T}\,(\beta_0+1) \right]} = \frac{\beta_0\, r_e}{\left[1 + j\,\frac{\omega}{\omega_{c\beta}} \right]} \qquad \cdots \quad (7.13.2)$$

Since

$$\frac{1}{1-\alpha_0} = (\beta_0 + 1) \simeq \beta_0,$$

this impedance is a parallel combination of a resistance $\beta_0 r_e$ and a capacitance $C = \dfrac{1}{\omega_T\, r_e}$.

For the component

$$i_1\, (y_{21e}/y_{11e}) = \beta\, i_b = \frac{\beta_0\, i_b}{(1 + j\ \omega/\omega_{c\beta})} \qquad \cdots \quad (7.13.3)$$

$\because$ i_1 is the input current, *i.e.*, base current in C.E. configuration. While the component $(y_{11e}/\Delta y_e)$ can be found out using equation (7.10.1) as follows :

The admittance $\dfrac{\Delta y_e}{y_{11e}} = \left[y_{22e} - y_{12e} \left(\dfrac{y_{21e}}{y_{11e}} \right) \right] \qquad \cdots \quad (7.13.4)$

Substitutingthe values for y_{22e}, y_{12e} and $\left(\dfrac{y_{21e}}{y_{11e}} \right)$ from equation (7.10.1) and (7.13.3) we have

$$\frac{\Delta y_e}{y_{11e}} = j\omega\ C_c \left(1 + \frac{\beta_0}{1 + j\omega/\omega_{c\beta}} \right) \qquad \cdots \quad (7.13.5)$$

This is sum of two admittances, one due to a capacitance C_e, the other due to $j\omega\ C_c\ \beta_0 \big/ \left(1 + j\, \frac{\omega}{\omega_{c\beta}} \right)$. The later admittance is due to an impedance. While this impedance is a series combination of resistance (Cr_e/C_c) and capacitance $\beta_0 C_c$. Therefore, the admittance is due to a capacitance C_c in parallel with an impedance formed of a resistance $\dfrac{Cr_e}{C_c}$ and a capacitance $\beta_0 C_c$ in series. For the range of frequencies $\omega \gg \omega_{c\beta}$, the component $\frac{y_{12e}}{y_{11e}} \cdot v_2$ becomes $\frac{C_c}{C} \cdot v_2$. Hence using these components and placing the base distributed resistance r_b in series, the h.f. equivalent circuit can be drawn as in Fig. (7.13.1).

For considering the components responsible for high frequency performance of a transistor, this circuit can be further simplified ; if we remember for very high frequencies $\omega \gg \omega_{c\beta}$ and higher, the first term in the denominator of eqn. (7.13.3) drops out, the component $(1/y_{11e})$ becomes capacitive and practically a short circuit. Lastly, ignoring the feedback which is small, for very high frequencies the equivalent circuit reduces approximately to that shown in Fig. (7.13.2).

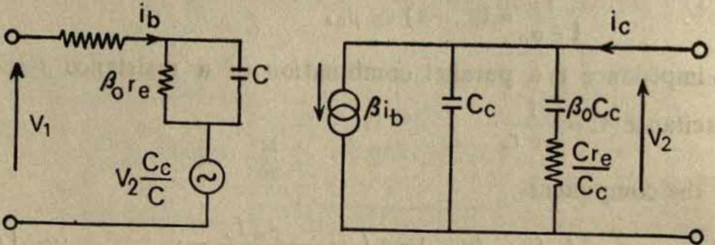


Fig. 7.13.1. H.F. equivalent circuit of transistor in C.E. configuration

7.14. Maximum frequency of oscillation f_m

The maximum frequency of oscillation is a frequency when the unilateral matched power gain falls to unity. For output matching one would need a variable inductor of value $(1/\omega^2 C_c)$ to tune out the capacitance C_c and an output load resistance of value (Cr_e/C_c), when the circuit of Fig. (7.13.2) becomes with this sort of load as shown in fig. (7.14.1). The power input to the transistor is $P_{in} =$

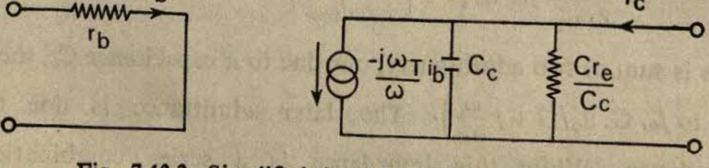


Fig.. 7.13.2 Simplified circuit for still higher frequency

$i_b{}^2 r_b$ while the matched power output is $\left(\frac{1}{2}\frac{\omega_T}{\omega}\, i_b\right)^2 \cdot \frac{Cr_e}{C_c}$ where $Cr_e = 1/\omega_T$. At very high frequencies the power gain becomes

$$G_{HF} = \left[\frac{\omega_T}{4 r_b\, C_c\, \omega^2}\right] \qquad \cdots \quad (7.14.1)$$

Note that this power gain falls at the rate of 6 dB/octave. This power gain becomes unity at a frequency, known as the maximum frequency of oscillation

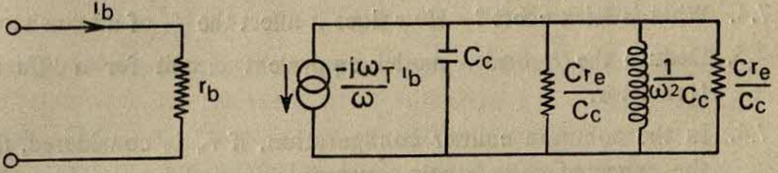


Fig. 7.14.1. Matched, output capacitance C_c tuned out, circuit

$$f_m = \left[\frac{f_T}{8\pi\, r_b\, C_c}\right]^{1/2} \qquad \cdots \quad (7.14.2)$$

This frequency is another transistor frequency of merit, dependent on the construction of the transistor which, in turn, decides the values of r_b, C_c and f_T. Lower r_b and C_c are, and higher f_T is, higher is the f_m also.

As an exercise the student can find for high frequencies, the output impedance in CE configuration. Hence show that for $\omega \leqslant \omega_{c\beta}$, the output impedance is a resistance given by the parallel combination of $\left[\frac{\alpha_{oi}\, Kr_e}{\alpha_o}\right]$ and $\left[\frac{Kr_e}{\alpha_o} \cdot \frac{\beta_{oi}}{\beta_o}\right]$.

References

1. *Fundamentals of Semiconductor Devices* : J. Lindmayer & C. Y. Wrigley, D. Van Nostrand Co. Inc. (1965).
2. *An Introduction to Semiconductor Electronics* : R. P. Nanabati, Mcgraw Hill Book Co. Inc. (1963).

Problems

7.1. Calculate the maximum low frequency power gain in CB configuration if the base resistance r_b is included in the calculation of equation (7.8.6).

7.2. Show that at f_T the transistor α has a real part equal to 0.5.

7.3. At $\omega = 0{\cdot}1\ \omega_T$, find the magnitude of β and its phase with respect to its low frequency value.

7.4. What is Kirk effect ? How does it affect the f_T of a transistor ?

7.5. Deduce the hybrid π model equivalent circuit for a diffusion transistor.

7.6. In the common emitter configuration, if r_b is considered, find the expected voltage gain, current gain and power gain when such stages are connected in tandem.

7.7. Find the input output matched voltage gain, current gain and power gain in CE configuration if r_b is not considered.

7.8. Argue to establish that the diode characteristics of the emitter base junction is much more steeper when the base collector are shorted than in the open collector configuration.

7.9. Find the f_T of a transistor having base width 2 microns, base region diffusion constant $D_n = 30$ cm^2/sec, base doping exponential, the exponent being $2\theta = 2{\cdot}3$, at a current of (a) 1 mA, (b) 5 mA, (c) 10 mA. Assume $C_e = 30$ P$_f$, $C_c = 2$ Pf and $r_b = 50$ Ohms.

7.10. Find f_{max} of the transistor of problem (7.9) above at 5 mA.

7.11. Find the low frequency power gain of the transistor of problem (7.12) ; hence voltage gain in CB configuration at 5 mA.

7.12. Find the low frequency power gain $\times f_{max}$ product of the transistor in problem (7.11).

7.13. Find the H.F. power gain at $f = m f_{max}$ of this transistor at 5 mA, where m is a fraction.

7.14. Find the low frequency output impedance of the transistor above in (a) CE configuration, (b) in CB configuration.

7.15. Draw the more accurate equivalent circuit of the junction transistor as indicated in article (7.14).

Chapter 8

TRANSIENT RESPONSE OF BIPOLAR TRANSISTOR

The ON resistance of a transistor is only a few Ohms whereas the OFF resistance can be easily of the order of a few hundered kilo-Ohms. Thus the transistor is an efficient switch. In switching applications, we are interested in the speed of the switch. Though the overall switching speed would also depend on the stray capacitances in the circuit, still it is worth knowing the maximum switching performance attainable if the stray capacitances were nil. In such ideal cases, the particular device speed capability will dictate the highest attainable speed.

8.1 Transient times

When a transistor is driven from OFF to ON and ON to OFF state by base current drive as shown in Fig. 8.1.1b, the collector current wave form, as in Fig. 8.1.1c, is obtained.

Though the base current step has occurred at t_{0+}, the collector current does not rise till a delay time (t_{a_1}), has passed. This time will be called the turn on delay and arises due to the passive input capacitance charging of the transistor from the OFF state to the cut-in state.

After this, the current in collector terminal rises and aims towards the final value. But it is difficult to recognise exactly where it started to rise from and is rather slow towards the end of its excursion. So the rise time t_r of the current waveform is defined as the time to rise up to 90% of its final value from its 10% value.

On the other hand, when the base is driven to turn the standing collector current off, between the time of application of the off drive to the time when the collector current starts decreasing, there is a time delay. This delay time to called the storage delay (t_s).

Durring this delay time, no appreciable change in output current waveform is observed.

After the expiry of the storage delay the collector current falls towards the zero value. The time that elapses from the end of storage time to the fall to 10% value of the collector current is called the fall time t_f.

$$\left.\begin{array}{l}\text{The turn on time of the transistor is } t_{\text{on}} = t_{d1} + t_r \\ \text{Similarly, the turn-off time of the transistor is } t_{\text{off}} = t_s + t_f\end{array}\right\} \quad \cdots \quad (8.1.1)$$

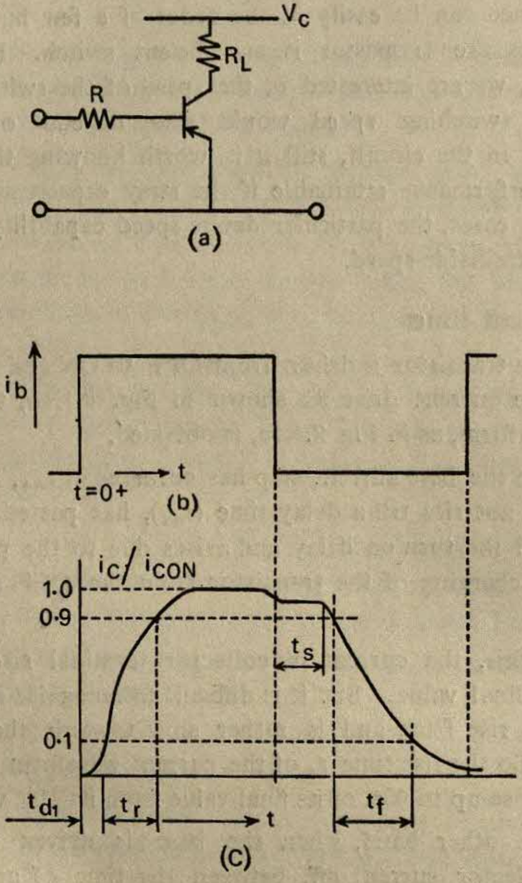


Fig. 8.1.1. (a) Transistor switch, (b) Base current wave form and (c) Normalised collector current wave form vs. time

8.2 Computational models

The methods that are used for the determination of ON-OFF times of the transistor are of three types. The equivalent circuit method, Ebers Moll model and the charge control method. In the equivalent circuit method one uses the small signal equivalent circuit and forward and reverse current transport parameters to calculate the switching times (cf. ref. 1). In the Ebers Moll model, use is made of equation of the form (6.4.1) and (6.4.2) to describe the transistor where the constant *b*'s are functions of frequency through forward and reverse α's (cf. ref. 3). Lastly, in the charge control model instead of the current, the total bulk charges ; those are

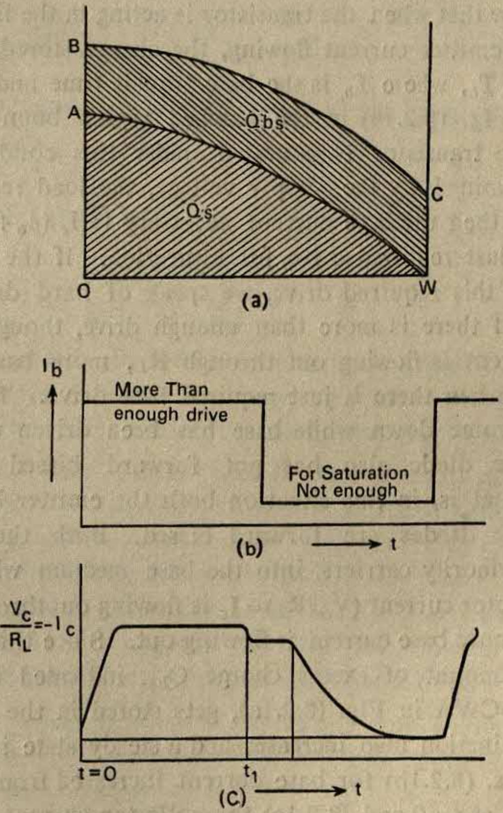


Fig. 8.2.1. (a) Shows base region charge, (b) Base current for over drive and insufficient drive, (c) Corresponding collector current

introduced by the signal voltage, are used to represent the signal and the information content therein. On the face of this it might appear that this would be an inaccurate method. Surprisingly enough, properly determined parameters can sufficiently accurately predict the transistor performance under all conditions. The conditions that we have in mind are : (*i*) the cut-off condition, when the base emitter junction is reverse biased and the collector current is cut-off, (*ii*) the active region, when the transistor is acting in its normally used amplifying mode and (*iii*) the saturation or storage region, that is, the region of operation when the transistor is expelling excess charges, that may have been stored earlier. Fig. (8.2.1) depicts this situation.

We know that when the transistor is acting in the linear region, for a certain emitter current flowing, the charge stored in the base region $Q_s = I_e T_b$, where T_b is the base transit time and this charge is shown in Fig. (8.2.1a) in the hatched region bounded by area AWO. If the transistor is saturated under this condition, which will be determined by the supply voltage, the load resistance and transistor β_0, then the base current necessary is $I_e/(\beta_0+1) = I_b$, will be called the just required drive, for saturation. If the base current is more than this required drive, we speak of hard drive or hard saturation. If there is more than enough drive, though the same collector current is flowing out through R_L, more base current is flowing than when there is just required base drive. The collector voltage has come down while base has been driven up thus the collector base diode also has got forward biased under this situation. That is, in this situation both the emitter base and the collector base diodes are forward biased. Both the diodes are pushing in minority carriers into the base medium while the old value of collector current $(V_{cc}/R_L) = I_c$ is flowing out through the load resistor and more base current is flowing out. Since this continues, some more amount of excess charge Q_{bs}, indicated by the area enclosed by BCWA in Fig. (8.2.1a), gets stored in the base region, hence recombination also increases and a steady state is reached as shown in Figs. (8.2.1b) for base current increased from not enough for saturation at $t = 0$ and (8.2.1c) for collector current. When the off drive or else a drive not enough for saturation is applied, until all the excess charge Q_{bs} has been expelled out, the transistor is incapable of entering in the active linear region, where the current flowing is described by $Q_s = I_e T_b$. The time that is needed to expel this excess charge, is the storage time, that we have referred to earlier in article (8.1).

In the following we shall use the charge control parameter description and use some of them conveniently to find the switching times ; combined with little bits of equivalent circuit approach as may be convenient for a particular purpose.

8.3 Charge control equations

The starting point of the charge control ideas come from firstly the continuity equation

$$\nabla . J_n = -\nabla . J_p = q\frac{\delta \hat{p}}{\delta t} + \frac{q\hat{p}}{T_p} \qquad \cdots \quad (8.3.1)$$

which may be integrated over the volume of the base and Gauss's theorem used to give

$$-\int_{\text{base}} \nabla . J_p \, dV = -\oint_s J_p . \, ds = \int_{\text{base}} \left(q\frac{\delta \hat{p}}{\delta t} + \frac{q\hat{p}}{T_p}\right) . \, dV \qquad \cdots \quad (8.3.2)$$

$$i.e. \quad -i_p = \frac{\delta Q}{\delta t} + \frac{Q}{T_p} \qquad \cdots \quad (8.3.3)$$

Equation (8.3.3) states that the hole current going out of the base surface plus the rate of hole charge loss due to recombination accounts for the total rate of minority charge loss from the base.

Secondly, the time constant used in the continuity equation is the life time but it is not the base alone, the extrinsic parameters as well determine time constants which are also responsible for transistors complete performance in any mode. Over and above, the frequency performance ; hence the involved time constants are in general different when compared between the forward and the backward (emitter and collector interchanged) mode of its working.

Thirdly that, whatever surfaces the charges cross within the bulk, there are three electrodes in terms of the currents of which, we need to formulate the charge flows.

To resolve these we note that equation (8.3.3) is a linear equation and it would not be improper to speak of superposition of currents and so addition of recombination share of current terms due to different time constants that may be necessarily proposed to describe the complete performance of the transistor. Obviously, the question would arise how many such time constants would be necessary ?

We have seen that a certain minority charge, introduced in base by emitter junction, calls for equal amount of majority carrier charge from base terminal and that the base and emitter currents are of different values. Hence for injection of the same magnitude of charge, base and emitter terminal descriptions need two different time constants in the forward mode of working of the transistor. Obviously, then if the transistor works in the reverse mode, that is the emitter and collector interchanged, two more time constants are necessary for description of charges of this mode. Thus a minimum of four time constants should be enough to describe charge storage in all modes of the transistors working.

From the above discussion it is clear that any terminal current can be written as the sum of currents due to rate of change of charges and the sum of currents due to recombination losses. With this in mind, we use here four time constants to write the terminal current equations of the transistor. For this we use here a P-N-P transistor whose both junctions are forward biased as shown in Fig. (8.3.1). In this the currents are written in terms of hole charge (positive) in the base and the base side of the junctions. For the base current, therefore, which is due to electrons entering, implies holes leaving the base terminal, we can write

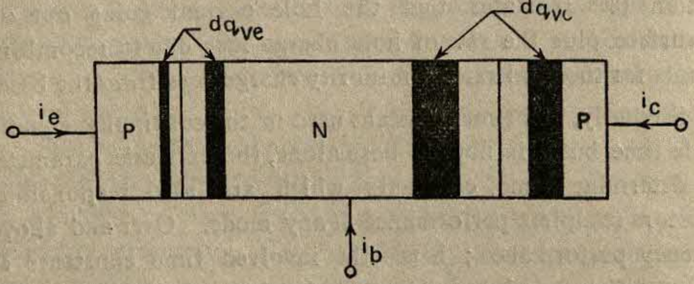


Fig. 8.3.1. Both junction forward biased transistor defining incremental charges

$$i_b = -\frac{q_f}{T_{bf}} - \frac{q_r}{T_{br}} - \frac{d}{dt}(q_r + q_f + q_{vc} + q_{ve}) \quad \cdots \quad (8.3.4)$$

where q_f is the forward hole charge injected into the base by the emitter with a time constant T_f, when an equal electron charge comes from the base terminal, T_{bf} being the base time constant responsible for this injection by base terminal. Similarly, q_r and T_{br} refer to the forward biased collector base junction injected hole charge and base terminal time constant involved respectively. While $(-dq_{ve})$ and $(-dq_{vc})$ are the electron charges necessary to cover up immobile exposed donor charges of the transition region on the base side for change of forward bias voltages dV_e and dV_c of the emitter base and collector base diodes.

Similarly, the other two equations may be written as

$$i_e = \frac{q_f}{T_f} + \frac{q_f}{T_{bf}} - \frac{q_r}{T_r} + \frac{d}{dt}(q_f + q_{ve}) \quad \cdots \quad (8.3.5)$$

$$i_c = \frac{q_r}{T_r} + \frac{q_r}{T_{br}} - \frac{q_f}{T_f} + \frac{d}{dt}(q_r + q_{vc}) \quad \cdots \quad (8.3.6)$$

assuming the forward bias diodes on either side are injecting q_f and q_r hole charges with respective emission time constants T_f and T_r respectively while injected hole charges (dq_{ve}) and (dq_{vc}) are covering up bare acceptor charges on each emitter side. In writing equations (8.3.4) to (8.3.6) we need to remember that Kirchoff's law is to be satisfied *i.e.*, sum of all terms on the R.H.S. of the equations (8.3.4) to (8.3.6) must be zero and that in no way equation (8.3.4) should be changed since it is a statement of the operation of the base region. Additionally, because of similarity of the roles of both the emitters, the terms $(+q_f/T_{bf})$ should be put in equation (8.3.5) and $(+q_r/T_{br})$ should be put in (8.3.6). Lastly, we see that we need to put $(-q_f/T_f)$ in (8.3.6) and $(-q_r/T_r)$ in (8.3.5) since (8.3.4) cannot be changed in any way.

Thus, we see that it has been possible to set up a set of general equations describing the terminal currents of a transistor in terms of charges and four time constants, which will embrace all working modes of the transistor. For different modes of operation the above set of equations may be written as :

For cut-off region :

In this region both junctions are reverse biased hence putting $q_r = 0$ and $q_f = 0$ in (8.3.4) to (8.3.6) we have

$$\left.\begin{aligned} i_b &= -\frac{d}{dt}(q_{ve} + q_{vc}) \\ i_c &= \frac{d\,q_{vc}}{dt} \\ i_e &= \frac{d\,q_{ve}}{dt} \end{aligned}\right\} \quad \cdots \quad (8.3.7)$$

For active region :

In the active region of operation of the transistor, the collector is reverse biased and the emitter is forward biased ; hence $q_r = 0$ put in equations (8.3.4) – (8.3.6) gives

$$\left.\begin{aligned} i_b &= -\frac{q_f}{T_{bf}} - \frac{d}{dt}(q_f + q_{vc} + q_{ve}) \\ i_e &= \frac{q_f}{T_f} + \frac{q_f}{T_{bf}} + \frac{d}{dt}(q_f + q_{ve}) \\ i_c &= -\frac{q_f}{T_f} + \frac{d}{dt}(q_{vc}) \end{aligned}\right\} \quad \cdots \quad (8.3.8)$$

For saturation region :

In the saturation region both V_{be} and V_{bc} are nearly constant so q_{ve} and q_{vc} also can be regarded to be constants ; when we have from (8.3.4) to (8.3.6)

$$\left.\begin{aligned} i_b &= -\frac{q_f}{T_{bf}} - \frac{q_r}{T_{br}} - \frac{d}{dt}(q_f + q_r) \\ i_e &= \frac{q_f}{T_f} + \frac{q_f}{T_{bf}} - \frac{q_r}{T_r} + \frac{d}{dt}(q_f) \\ &\text{and} \\ i_c &= \frac{q_r}{T_r} + \frac{q_r}{T_{br}} - \frac{q_f}{T_f} + \frac{d}{dt}(q_r) \end{aligned}\right\} \quad \cdots \quad (8.3.9)$$

An extensive analysis shows (ref. 1.3) that for the saturation mode there are two natural modes of transients that are excited by base drive of which one is much faster than the other while the faster one has smaller amplitude compared to the slower mode. Hence in the case of equations (8.3.9) we look for a simpler statement of equations utilising only one time constant obviously corresponding to the larger time constant one. This leads to the simpler way of looking at the saturation mode in the working range of the transistor when the excess charge (q_{bs}) is being expelled with a time constant T_s and the equations describing this can be written as

$$\left.\begin{aligned} & -i_{bs} = \left(\frac{q_s}{T_s}\right) + \frac{dq_s}{dt} \\ \text{where} \quad & \\ & q_s = q_f + q_r - I_c\,T_f \\ \text{and} \quad & \\ & i_{bs} = i_b - \frac{I_c}{\beta_f} \end{aligned}\right\} \quad \cdots \quad (8.3.10)$$

Note that i_{bs} is the excess part of the base current that is driving out the excess base charge q_{bs} only.

8.4 Measurement of the Switching parameters

In this section we shall study a method for the determination of the time constants T_f and T_{bf} while the other two T_r and T_{br} can be found out by the same method with the roll of emitter and collector interchanged. Taking the active region equations, if we confine our attention only to the steady parts of the collector waveform, when the transistor is being driven as shown in Fig. (8.4.1) by a train of square wave pulses, then we can write

$$i_b = \frac{-q_f}{T_{bf}} \quad \cdots \quad (8.4.1)$$

$$i_c = \frac{-q_f}{T_f} \quad \cdots \quad (8.4.2)$$

We assume that a measured steady, initial d.c. collector current I_{Ci} is flowing through the transistor on which the driving base current change ($\triangle\, i_b$) has caused a change ($\triangle\, i_c$), then from equations (8.4.1) and (8.4.2) we can write

$$T_{bf} = \frac{-\triangle q_f}{\triangle i_b} \text{ and } T_f = \frac{-\triangle q_f}{\triangle i_c} \quad \cdots \quad (8.4.3)$$

Resistance R_1 in parallel with R_2 is assumed large enough for their current shunting effect to be ignored. Pulse train derived from the same source *i.e.*, of same P.R.F. and pulse width and occurring simultaneously but each of separately variable heights (V_1, V_2 and V_3) are applied to the three terminals as shown in Fig. (8.4.1). The resistance R at the base decides the steady state

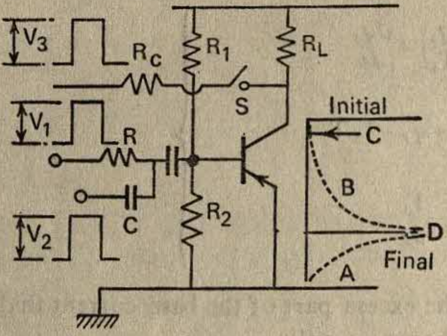


Fig. 8.4.1. Switching time measuring circuit

collector current by the drive $\frac{V_1}{R} = \Delta i_b$. The condenser gives an instantaneous charge input to the base medium of value CV_2, this charge will instantaneously change the current in the transistor but whether this value will be maintained or not, would be decided by the Δi_c step, determined by V_1. If V_1 —produced Δi_c is less than the instantaneous rise of current caused by V_2 then the collector voltage would change along a curve like A shown in the inset of Fig. (8.4.1). If, on the other hand, V_2 hence CV_2 is small enough, that the instantaneous collector current rise is lesser than the steady state Δi_c change given by V_1 and R, then the collector voltage waveform will fall along a curve like B from a point like C towards the level D. Therefore, by adjusting V_1 or V_2 or both it is possible to have a square shape drop of collector voltage waveform as shown by curve D.

From equations (8.4.3)

$$\frac{\Delta i_c}{\Delta i_b} = \beta_f = \frac{T_{bf}}{T_f} \qquad \cdots \quad (8.4.4)$$

The charge given by condenser C into the base of the transistor distributes itself as shown by

$$(-CV_2) = \Delta q_f + \Delta q_{vc} + \Delta q_{ve} \qquad \cdots \quad (8.4.5)$$

From (8.4.5) and (8.4.3)

$$T_{bf}=\frac{-\Delta q_F}{\Delta i_b}=\frac{-(\Delta q_F)R}{V_1}=\frac{R.CV_2}{V_1}\;\frac{\Delta q_f}{\Delta q_f+\Delta q_{vc}+\Delta q_{ve}} \quad \cdots \quad (8.4.6)$$

$$\text{and } T_f=\frac{-\Delta q_f}{\Delta i_c}=\frac{CV_2\,\Delta q_F}{\Delta i_c\,(\Delta q_F+\Delta q_{vc}+\Delta q_{ve})} \quad \cdots \quad (8.4.7)$$

From equation (8.4.6) we see, if we determine T_{bf} by $\left(\frac{V_2}{V_1}\cdot RC\right)$, then the estimated T_{bf} is more than its actual value. Hence in the experiment, our aim is now to make Δq_{vc} and Δq_{ve} equal to or as near to zero as possible. If we can do so, then T_{bf} estimated by $\left(\frac{V_2}{V_1}\cdot RC\right)$ would be the required T_{bf}.

After the adjustments of V_1 and V_2 has been made to achieve square top collector voltage waveform as mentioned above, the switch S is closed and the voltage V_3 can be adjusted to push up the collector voltage such that V_{cb} do not change ; this would ensure $\Delta q_{vc}=0$. But this would need either a differential oscilloscope or an oscilloscope whose both input terminals are floating. If these are not available the oscilloscope can be connected between collector to ground and ΔV_{ce} is made zero by adjusting V_3. Since ΔV_{be} is of the order of a few milivolts, therefore, making $\Delta V_{ce}=0$ is enough of a good approximation of $\Delta V_{cb}=0$, provided prior to closing the switch S, ΔV_{cb} was large compared to a few milivolts. With this adjustment of V_3, therefore, we have achieved practically $\Delta q_{vc}\simeq 0$.

It should be noted priod to applying the pulses V_1, V_2 or V_3 we have measured I_{ci}. Therefore, after adjustment of V_3 we can estimate T_{bf} by $\left(\frac{RC\,V_2}{V_1}\right)$, and T_f from (8.4.7), with $\Delta i_c=(V_3/R_c)$ by $\left(CR_c\,\frac{V_2}{V_3}\right)$ for each value of I_{ci} selected. But since $\Delta q_{ve}=0$, the values of T_{bf} or T_f thus obtained would be more than their actual values. But as more and more increased I_{ci} values are selected for the same (Δi_c), less and lesser are the ΔV_{be} required, because of the near exponential nature of the input characteristic of the transistor. Hence choice of increased I_{ci} values lead to lessening of the charge (Δq_{ve}). So, if we plot T_f & T_{bf} estimated by

$$T_f=\frac{CR_c\,V_2}{V_3}\quad\text{and } T_{bf}=\frac{CR\,V_2}{V_1} \quad \cdots \quad (8.4.8)$$

for different initially measured I_{cs} values, we get curves as shown in Fig. (8.4.2) which would asymptotically approach to their exact values for large I_{ci}. Therefore, the experimental steps may be summarised as follows :

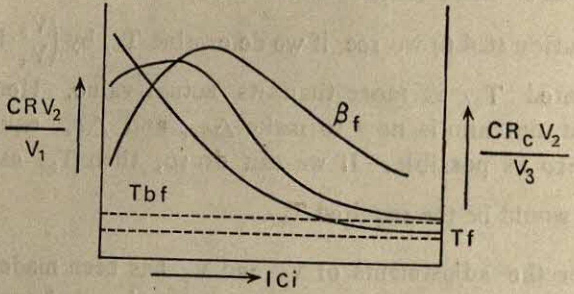


Fig. 8.4.2. Plot of CRV_2/V_1, CR_cV_2/V_3 and their ratio β_f

(*i*) With V_1 and V_2 adjusted to zero and switch S open, adjust and measure I_{ci}. Then (*ii*) adjust V_1 and V_2 to get square type collector voltage of a few volts. (*iii*) close switch S and adjust V_3 to make $V_{ce}=0$ seen on an oscilloscope connected between ground and collector, then (*iv*) measure the heights representing V_1, V_2 and V_3 in the same scale of the oscilloscope since we are interested in their ratio, the absolute value of the voltages are not needed, (*v*) open switch S and adjust $V_1=V_2=0$, readjust I_{ci} to a larger measured value and repeat steps (*i*) to (*v*). Lastly, plot I_{ci} versus T_f and I_{ci} versus T_{bf} computed by equation (8.4.8) and determine the asymptotic values.

8.5 Physical meanings of the time constants

In order to get an insight into the physical origin of the time constants T_f and T_{bf}, we start with approximate active region equations obtained from (8.3.8) by ignoring $\left(\frac{d}{dt}q_{vo}\right)$ in i_b and $\left(\frac{d}{dt}q_{vc}\right)$ in i_c.

These two equations are

$$i_b=-\frac{q_f}{T_{bf}}-\frac{d}{dt}(q_f+q_{vc}) \quad \cdots \quad (8.5.1)$$

$$i_c = -\frac{q_f}{T_f} \qquad \cdots \quad (8.5.2)$$

The approximations mentioned just now are not as drastic as may seem at first sight.

Experimental results suggest that these assumptions are usable. Also, though the collector base capacitance changes widely during the transients, an average value of this parameter can be found out for typical junctions (cf. ref. 1). In our calculations we shall use this to be C_c and regard it as a constant.

With the above mentioned reservations we calculate

$$\frac{d}{dt}(q_{vc}) = \frac{d}{dt} C_c (V_{ce}) = \frac{d}{dt}(V_c - i_c R_L) = -C_c R_L \frac{di_c}{dt} \qquad \cdots \quad (8.5.3)$$

Equation (8.5.3) with (8.5.2) and (8.4.4) gives

$$\frac{d\, q_{vc}}{dt} = \frac{C_c R_L}{T_{bf}} \beta_f \frac{d\, q_f}{dt} \qquad \cdots \quad (8.5.4)$$

Substituting (8.5.4) in (8.5.1) gives for i_b

$$i_b = -\frac{q_f}{T_{bf}} - \left(1 + \frac{\beta_f C_c R_L}{T_{bf}}\right) \frac{d\, q_f}{dt} \qquad \cdots \quad (8.5.5)$$

Thus we have got an approximate relation, the base current has, to the base stored charge in the active region of working of the transistor.

Let us now apply equation (8.5.5) to the output short circuit situation when the base current is of the form

$$i_b = i_{bo} \exp(j\omega t) \qquad \cdots \quad (8.5.6)$$

Since equation (8.5.6) is a linear differential equation, obviously q_f will also be of the form

$$q_f = q_{fo} \exp(j\omega t) \qquad \cdots \quad (8.5.7)$$

Substituting (8.5.6) and (8.5.7) in equation (8.5.5), we have for $R_L = 0$

$$q_{fo} = \frac{-i_{bo} T_{bf}}{1 + j\omega T_{bf}} \qquad \cdots \quad (8.5.8)$$

hence using equation (8.5.2) with equation (8.5.8,) we have

$$i_c = \frac{\beta_f i_{bo} \exp(j\omega t)}{1 + j\omega T_{bf}} = \frac{\beta_f i_b}{1 + j\omega T_{bf}} \qquad \cdots \quad (8.5.9)$$

Equation (8.5.9) shows that T_{bf} corresponds to $(1/\omega_{c\beta})$ of small signal theory of transistor but applicable to large signals. Since $T_{bf}=\beta_f T_f$ hence T_f correspondingly is the large signal counterpart of $(1/\omega_T)$ and so β_f is the large signal forward beta of the transistor.

8.6 Turn on delay

For the calculation of this delay time, we shall here study the equivalent circuit method. In this phase the transistor is being turned on from the cut-off state and for simplicity we assume the junction voltage in the cut-off state was V_2 volts. On application of a step voltage $(-V_1)$ at the input of the base terminal through a suitably large resistance, the base emitter junction voltage is being established to the cut-in value. Below this voltage the junction is not sufficiently biased to be counted in practice as conducting. (We shall take this as the just ON voltage of the junction.) During this delay period the current is not flowing since the junction voltage has not reached the value necessary for current flow ; so, during this time the transistor is passive. Hence we draw in Fig. (8.6.1a) the passive circuit part of Fig. (8.1.1a) to state the equivalent circuit applicable during this delay period. In this case, with change of junction voltages the capacitances also change ; but we assume here that they are constants and having suitably selected mean values (cf. ref. 1). C_e refers to the passive capacitance at the emitter base junction ; similarly, C_c refers to the capacitance for the collector base junction.

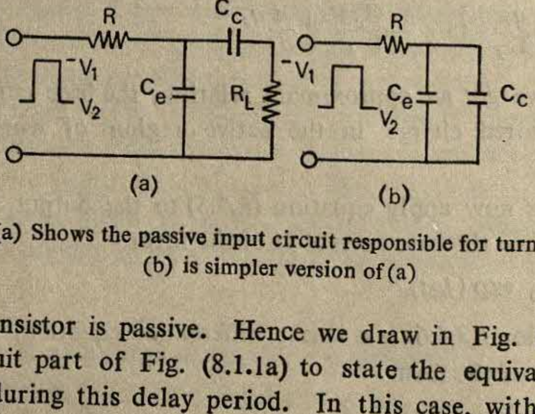


fig. 8.6.1. (a) Shows the passive input circuit responsible for turn on delay, (b) is simpler version of (a)

$(C_c R_L)$ time constant in practice being very small in high speed transistor application, the equivalent circuit can be reduced to as shown in Fig. (8.6.1b) by ignoring R_L. The charging of the emitter base voltage from a reverse bias to the cut-in voltage V_γ occurs with a time constant $T_\delta = R(C_e + C_c)$. Hence the charging equation may be written as

$$v(t) = V_2 - (V_2 + V_1)(1 - \exp - t/T_\delta) \quad \cdots \quad (8.6.1)$$

No transistor current flows as long as $|v(t)| < -V_\gamma$, assuming that the time needed is T_d from (8.6.1) we have

$$-V_\gamma = V_2 - (V_2 + V_1)(1 - \exp - T_d/T_\delta) \quad \cdots \quad (8.6.2)$$

From (8.6.2) we have the turn on delay given by

$$T_d = T_\delta \log_e \left(\frac{V_2 + V_1}{V_1 - V_\gamma}\right) \quad \cdots \quad (8.6.3)$$

Since V_γ is small hence we see, T_d decreases with increase of V_1 magnitude.

8.7 Transistor Rise Time

Case—I : Just required drive.

Firstly, we shall discuss the rise time of the transistor when the drive is just sufficient to saturate the transistor at the end of the transient period. For this purpose, we start from equation (8.5.5) which can be re-written as

$$-i_b \frac{T_{bf}}{T_{BF}} = \frac{q_f}{T_{BF}} + \frac{d}{dt} q_f \quad \cdots \quad (8.7.1)$$

$$\text{where } T_{BF} = T_{bf} + \beta_f C_c R_L \quad \cdots \quad (8.7.2)$$

It would be recognised that the second term in equation (8.7.2) arises from Miller effect of collector base capacitance.

Say, we have applied a step voltage $(-V_1)$ to put a step current drive of value $I_{bo} = \left(\frac{-V_1}{R}\right)$ at the base of the transistor, which is just sufficient to saturate the transistor. Then solving eqn. (8.7.1) for step base current, we have by using Laplace transform method with boundary condition $t = t_{0-}$; $q_f = 0$,

$$q_f(s) = \frac{-I_{bo} T_{bf}}{s(1 + sT_{BF})} \quad \cdots \quad (8.7.3)$$

From which we have

$$q_f(t) = -I_{bo} T_{bf} (1 - \exp - t/T_{BF}) \quad \cdots \quad (8.7.4)$$

Hence from (8.7.4) and (8.5.2)

$$i_c = I_{bo}\ \beta_f\,(1 - \exp - t/T_{BF}) \qquad \cdots \quad (8.7.5)$$

The collector current needs, for its excursion from 10% to 90% of full value ($I_{bo}\ \beta_f$), a time

$$T_r = 2.3\ T_{BF} - 0.1\ T_{BF} = 2.2\,(T_{bf} + \beta_f\, C_c\, R_L) \qquad \cdots \quad (8.7.6)$$

for the just required drive case.

Case—II : Over driven-transistor rise time

If I_{bi} is the step drive current where $|I_{bi}| > |I_{bo}|$ then from equation (8.7.1) with $q_f = 0$ at $t = 0-$, we can again find the stored charge variation as

$$q_f\,(t) = -\,I_{b1}\ T_{bf}\,(1 - \exp - t/T_{BF}) \qquad \cdots \quad (8.7.7)$$

and the collector current variation as

$$i_c = \frac{-q_f}{T_f} = I_{b1}\ \beta_f\,(1 - \exp - t_1/T_{BF}) \qquad \cdots \quad (8.7.8)$$

From (8.7.7) and (8.7.8) we conclude that when the stored charge in the base would be $I_{bo}\ T_{bf}$, the collector current will be ($I_{bo}\ \beta_f$) ; the collector current for just saturation would occur and the rise would be completed, hence equate

$$I_{bo}\ \beta_f = I_{b1}\ \beta_f \left(1 - \exp \frac{-t_r}{T_{BF}}\right) \qquad \cdots \quad (8.7.9)$$

to obtain for the rise time

$$t_r = T_{BF}\ \log_e \left(\frac{I_{b1}}{I_{b1} - I_{bo}}\right) \qquad \cdots \quad (8.7.10)$$

From (8.7.9), one can also find 10% to 90% of ($I_{bo}\ \beta_f$) excursion time. Remember once again that $T_{BF} = (T_{bf} + \beta_f\, C_c\, R_L)$. Equation (8.7.10) shows larger I_{b1} is ; lesser is the turn on time *i.e.*, harder drive brings the transistor faster to saturation. As a matter of fact, if $\frac{I_{bo}}{I_{b1}} << 1$, then (8.7.10) gives, expanding the logarithmic term and keeping only one term,

$$t_r = \left(\frac{T_{BF}\ I_{bo}}{I_{b1}}\right)$$

8.8 Saturation delay

When the transistor has been driven to hard saturation by current I_{b1}, the excess charge q_{bs}, stored in the base, is $(|I_{b1}| - |I_{bo}|)\,T_s$. Now saturation delay is the time taken to extract this excess stored charge. For the removal of this stored charge, we use equation (8.3.10) repeated here as

$$-i_{bs} = \frac{q_s}{T_s} + \frac{d\,q_s}{dt} \qquad \cdots \quad (8.8.1)$$

If the base drive voltage V_2 is applied such that a back bias for the emitter base junction is aimed at, then the drive current is $I_{b2} = V_2/R$, and the actual step input to the base for excess stored charge removal is $(|I_{b2}| + |I_{b1}|)$ while the excess stored charge is $(|I_{b1} - I_{bo}|)\,T_s$. Note that in a *p-n-p* transistor, the base current is flowing out of base terminal so I_b is negative. If this was borne in mind, the magnitude signs could be discarded with. But the magnitude sign has been used to avoid confusion. Hence in equation (8.8.1) using $(|I_{b2}| + |I_{b1}|)$ as the step input current and sloving for $q_s(t)$ with the boundary condition that at $t = 0-$, the stored charge was $(|I_{b1} - I_{bo}|)\,T_s$, we have,

$$\begin{aligned} q_s(t) &= (|I_{b1}| - |I_{bo}|)\,T_s \exp -t/T_s \\ &\qquad - (|I_{b2}| + |I_{b1}|)\,T_s(1 - \exp - t/T_s) \\ &= -(|I_{b2}| + |I_{b1}|)\,T_s + (2\,|I_{b1}| + |I_{b2}| - |I_{bo}|)\exp(-t/T_s) \end{aligned} \qquad \cdots \quad (8.8.2)$$

Note that at the base terminal a voltage V_2 has been applied with respect to ground to turn the transistor off; while equation (8.8.2) shows the total current change the transistor has been forced to aim at, is the sum $(|I_{b2}| + |I_{b1}|)$ which it should for applied back drive. The excess charge would be zero in the transistor after a time t_s of the application of drive voltage V_2 and is given from eqn. (8.8.2) as

$$t_s = T_s \log_e \left(\frac{|I_{b2}| + 2|I_{b1}| - |I_{bo}|}{|I_{b2}| + |I_{b1}|} \right) \qquad \cdots \quad (8.8.3)$$

Thus if saturation time constant T_s is known, then saturation delay may be calculated by equation (8.8.3). Note that greater I_{b2} is, lesser is t_s, the storage time.

8.9 Fall time with over-drive

By over-drive we mean that at the base terminal a voltage V_2 has been applied due to which the excess base charge has gone out; thereafter the drive is trying to back bias the emitter base junction. Hence, again we can solve equation (8.5.5) or (8.7.1) with equation (8.5.2) for $i_c(t)$ with the boundary values at the end of saturation delay $t = t_s$, (after the base turn-off drive was applied). At $t = t_s$, the collector current was $-\beta_f I_{bo}$ while at $t \to \infty$, $i_c \to +\beta_f I_{b2}$. In this we need to remember that active region base currents are negative. Since we know that the time constant involved is T_{BF}, given by equation (8.7.2) in this decay process, without solving the equation (8.7.1), we can write the collector current as

$$i_c(t) = -\beta_f I_{b2}\left(1 - \exp - \left(\frac{t - t_s}{T_{BF}}\right)\right) + \beta_f I_{bo} \exp - \frac{(t - t_s)}{T_{BF}} \quad \cdots \quad (8.9.1)$$

Hence the time passed in reduction of i_c to zero value since t_s, is from equation (8.9.1)

$$t_f = T_{BF} \log_e \frac{|I_{b2}| + |I_{bo}|}{|I_{b2}|} \quad \cdots \quad (8.9.2)$$

which is approximately the required fall time of article (8.1).

On the other hand, if the transistor was not earlier over-driven to saturation, rather was in just saturation with collector current $\beta_f I_{bo}$ flowing and the base drive have been reduced to zero, then we need to take a stored charge in the transistor at $t = 0-$ as $(-I_{bo} T_{bf})$ (cf. equation 8.7.4) and the forcing $i_b = 0$ to solve equation (8.7.1) and using equation (8.5.2) for $i_c(t)$ we have

$$i_c(t) = +\beta_f I_{bo} \exp(-t/T_{BF}) \quad \cdots \quad (8.9.3)$$

from which 10% to 90% fall time can be found as before as 2.2 T_{BF}.

References

1. *An Introduction to Semiconductor Electronics*: R. P. Nanabati, McGraw Hill Book Co. Inc. (1963).

2. *Physics of Semiconductor devices*: S. M. Sze, Wiley Eastern Ltd. (1979).

3. *Analysis and Design of Integrated Circuits*: McGraw Hill Book Co. (1967), p. 498.

Problems

8.1. Explain clearly the origin and meaning of the terms, (*a*) Turn on delay, (*b*) Rise time, (*c*) Saturation delay and (*d*) Fall time in a switching transistor in C.E. mode.

8.2. An *n-p-n* silicon transistor switch working from 9V supply has a large signal forward h_{FE} of 20. It is driven by voltage pulses of ± 5V through a 10 K. ohm resistor. Assuming average values of the base emitter capacitance as 3 Pf, collector base capacitance as 0.2 Pf, a load resistance of 1 K. ohm, and a forward large signal injection time constant as 2 *n* sec., find the total turn on time.

8.3. An *n-p-n* transistor, just at cut-off with base to emitter voltage nearly zero, operates with $V_{cc} = 12V$ and $R_L = 1$ K. ohm. A positive step input voltage is applied through a 10 K. ohms resistance to the base. The upper level of the input signal is 10V. It is found that the transistor reaches saturation in 0.3 micro secs. The transistor large signal $h_{FE} = 30$ and the average collector transition region capacitance is 0.3 Pf. Find the large signal f_T of the transistor.

8.4. A transistor base is switched from +2 volts to −2 volts and back. Calculate the turn on delay, turn on time and the saturation delay if the resistance in the base to driver is 10 K. ohms, the collector load is 2 K. ohms, $C_e = 1.5$ Pf and $C_c = 0.5$ Pf, $T_{BF} = 5 \times 10^{-9}$ secs., $h_{FE} = 50$ and $V_{cc} = 10$ volts and $T_s = 10^{-8}$ secs.

8.5. Calculate the capacitance needed to be put across the driver resistance 10 K. ohms to make the rise time in problem (8.4) above, − zero. Draw the collector voltage waveform and estimate the highest gating rate possible in this case. How can this rate be improved ?

8.6. Which of the following transistors having identical lifetime and resistivities of *n*- and *p*-regions has the highest speed of response (1) *n-p-n*, (ii) *p-n-p* of Ge, of Si or of GaAs ? And Why ?

8.7. What differences are there between a high frequency transistor and a high speed transistor ?

8.8. A transistor in deep saturation in CE mode giving the same output voltage as at the base. Explain why ?

8.9. A transistor switch has a capacitive load. Draw the load line expected on the output characteristics if the base is driven with steep edge voltages ±2 volts through a 10 K. ohm resistance and the transistor $\beta_F = 50$.

8.10. If in problem (8.9) above the load is inductive having a time constant lower than T_F, find the load line expected if the resistance of the inductance in parallel to the output conductance is sufficient to cause (*i*) saturation, (*ii*) non-saturation, for the drive current used.

8.11. Assuming that the collector base junction of an *n-p-n* transistor is a step junction and that the collector base junction voltage changes, in the inverter from a voltage $V + V_b$ to $0.5\ V_b$. Find the average collector base capacitance. Hence for $V_b = 0.5$ volts, $V = 20\ V_b$, $\varepsilon_r = 12$, $A = 400$ sq. microns and a doping level of 10^{15} for the base region and 10^{14} for the collector epitaxial layer, find the average collector junction capacitance to be used.

Chapter 9

JUNCTION FIFLD EFFECT TRANSISTOR

In this Chapter we shall study a device that is known as the Junction Field Effect Transistor abbreviated as JFET. For an understanding of the effect of field in semiconductor, we recall that the velocity field characteristics of semiconductors like Ge and Si are as given in fig. (2.11.4). It is to be noted from this figure that, for example, in *n*-type Si the velocity is proportional to the electric field till about 2 kv/cm while the velocity nearly saturates at about a field of $E_c = 20$ kv/cm. So, if we have an *n*-type semiconductor with two ohmic contacts as source S and drain D, on two n^+ doped regions on it and a voltage V_d is applied between them, then the current voltage characteristic is expected to be as shown in Fig. 9b diode (1). This is because the current is given by $I_d = Aq\,n\,v_{\text{drift}}$. At low voltages that is at low fields the drift velocity is proportional to the field E which again is proportional to V_d ; hence Ohm's law holds. At fields higher than 2 kv/cm in Si the constancy of mobility ceases and the current voltage characteristic does not follow Ohm's law. When the field has reached about 20 kv/cm, the drift velocity saturates so does the current at point P in curve marked diode 1.

If we have a *p*-doped region (gate G) in between the n^+-regions as shown in Fig. (9a), the saturation current will be reached a little earlier in voltage scale and that this current will be a little lower since the *n*-region (channel) thickness decreases by the *p*-region thickness and the accompanying depletion width of the *p-n* junction. This saturation point is shown by point Q in curve diode 2 in Fig. (9b).

If the ohmic contact to the *p*-region is connected to the source contact S and the voltage between S and D is increased from zero the current will follow Ohm's law at the low voltages and will again saturate corresponding to point R in Fig. (9b), but at a still lower voltage and lower current since the *p-n* junction is reverse biased and

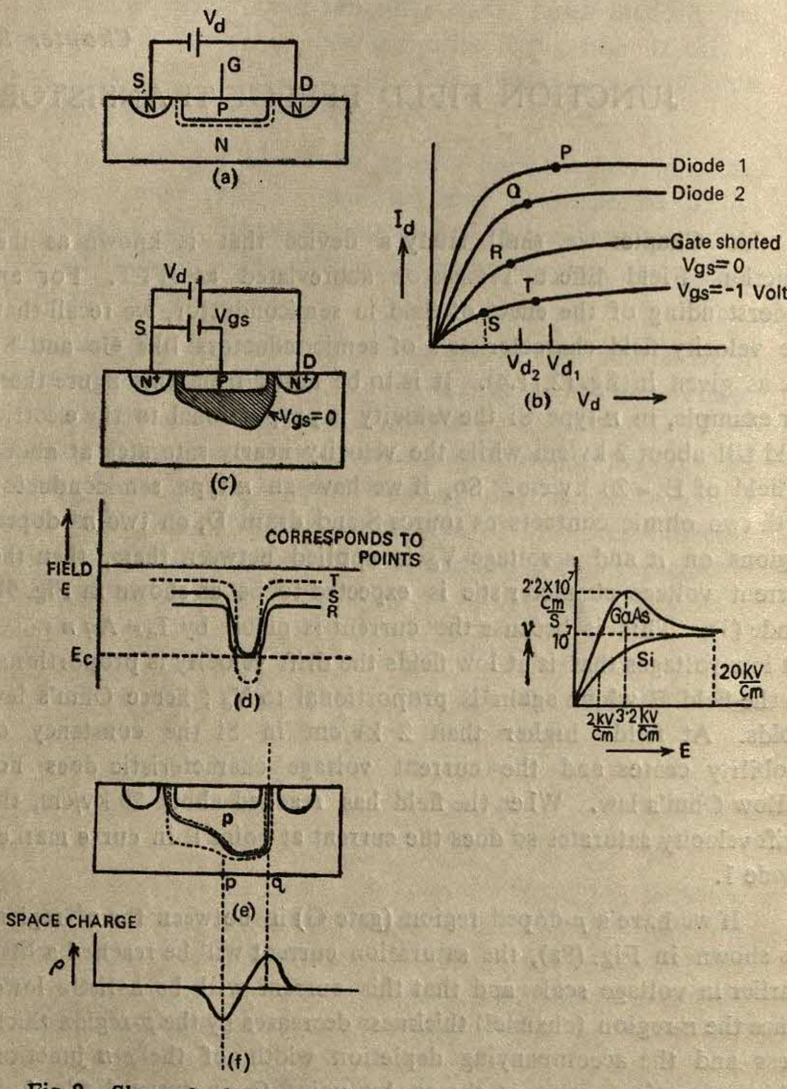


Fig. 9. Shows the depletion region field and space charge corrsponding to different points on the characteristics [Charles. A. Liechti Copyright © July 1976 IEEE]

this bias voltage increases from the source to the drain end of the junction. This causes the *p-n* junction depletion region width to increase leading to a decrease of channel thickness. Current saturation occurs when the field at the constricted point of the channel has reached the critical field E_c (~ 20 kv/cm in Si) as shown in Figs. (9c and 9d). When a negative bias voltage V_{gs} is applied, the *p-n* junction depletion region width still further increases as shown in Fig. (9e dotted line) and hence the current saturation occurs at a still lower V_d and a lower current and this saturation corresponds to point S in Fig. (9b). Increased width of the constricted region implies E_c is obtained in a wider region than at situation corresponding to R hence on either side of this region the field must be lesser, that accounts for a less current at S of Fig. (9b).

If the voltage V_d increases further than that corresponding to point S the field at the constricted part of the channel increases above E_c giving a wider high field region in the channel as *p* to *q* in Fig. (9e). This causes more voltage drop in this high field region leaving less voltage across the junction at any point between source end to point *p* (cf. Fig. 9d and 9e). This, in turn, widens the channel thickness on the left of *p* that accounts for slow increase of current with voltage after saturation at point S (cf. fig. 9b). The increased higher field region *pq* in the channel implies that the velocity is higher than outside this region. Hence positive space charge on the right and negative space charge on the left of the peak field point is expected in this region, as shown in Fig. (9f). This is because at the peak field point the velocity is highest, on the left lower ; hence at the peak field point electron carriers are removed faster than they arrive from the left. The field due to this space charge accounts for the drop of increased applied voltage in this region. Thus after saturation the current slowly increases with further increase of V_d and this space charge region accounts for the finite slope of the output characteristics after saturation. Note that the depletion increases in extension more than the right-hand edge of the gate.

Thus we see that the JFET consists of a semiconducting channel whose thickness can be varied by widening the depletion region of the *p-n* junction and the depletion region widening is the effect of the applied field through the applied voltage.

In the following article, we shall simplify the picture presented above such that analytical current voltage expressions can be found out easily.

9.1 Junction FET d.c. characteristics

In the junction field effect transistor (JFET) majority carriers flow through the bulk semicouductor which have built in *p-n* junction or junctions. These *p-n* junctions are used to control the resistance of the path of majority carriers by backward biasing as shown in Fig. (9.1.1). Fig. (9.1.1a) shows a double gate and (9.1.1b) shows a signal gate Junction FET. The majority carriers find a narrowed down cannel to flow through. The *p-n* junctions are assumed to be abrupt junctions—the *p*-regions are heavily doped so that the depletion regions mostly extend into the *n*-regions. The majority carriers are injected by the ohmic source contact S and drained out by the ohmic drain contact D. The gate electrode G makes ohmic contact to the *p*-type doping. The cross section of the channel is controlled by varying the source to gate voltage V_g or the source to drain voltage V_d or jointly by both. Note that when both V_g and V_d are applied maximum reverse bias of *p-n* junctions occur at the drain ends at $x = L$ and this reverse bias is $(V_g + V_d)$ while the depletion region voltage is $(V_g + V_d + V_b)$ where V_b is the built in voltage of the junctions. On the other hand, at $x = 0$ the voltage across the depletion region is $(V_g + V_b)$. The junction voltage at any point x is $(V_g + V_b + V)$ where V is the magnitude of the voltage on the channel side of the depletion region boundary measured with respect to the source, where $x = 0$ and $V = 0$. Referring to Fig. (9.1.1a) the channel thickness y at x can be written, assuming symmetrical doping, as

$$2y = 2a\left(1 - \frac{t}{a}\right) \qquad \cdots \quad (9.1.1)$$

where t is the depletion region width and $2a$ is the width of the *n*-region. With increased applied voltages the depletion regions extend to have no channel. This occurs at the drain end first. The voltage across the depletion region at this situation is a particular device constant and is called the pinch-off voltage. If source drain voltage V_d is not applied, then a greater gate voltage would be needed to pinch off the channel and this would occur throughout from $x = 0$ to $x = L$. If this junction voltage is V_p then $V_p = \frac{qN}{2\varepsilon\varepsilon_0} \cdot a^2$; hence

$$t = a \cdot \sqrt{\frac{V_g + V_b + V}{V_p}} \qquad \cdots \quad (9.1.2)$$

Since the drain current flows in an wedge-shaped region, therefore in addition to field opposite to the x direction, there will also have

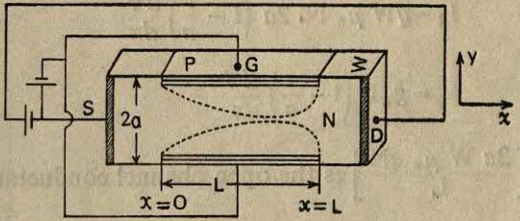


Fig. 9.1.1a. Double gate JFET—both gates connected to source

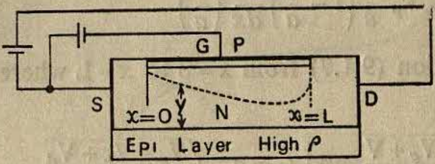


Fig. 9.1.1b. Single gate connected source

field in the y direction trying to push the carriers towards the axis. That is the equipotential lines will not be plane parallel but to some extent curved parallel. Hence $\frac{\delta^2 V^{(x\ y)}}{\delta x^2}$ is not zero in the channel *i.e.*, there would be space charges in the channel which is in contradiction to the abrupt junction assumption. If, of course, the space charge indicated by $\varepsilon\varepsilon_0 \frac{d^2V}{dx^2}$ is very much less than (qN) the mobile charge density, then the y directional field would be small compared to the x directional field. In that case, the junction boundary slope would be small, that is, the channel width varies slowly and we speak of gradual channel approximation.

To find the current as a function of the applied voltages, we can write for the current that flows through the channel

$$I_d = -\int_{-y}^{y} J_x W \, dy \qquad \cdots \quad (9.1.3)$$

where W is the channel width perpendicular to the x and y directions (*c.f.* Fig. 9.1.1). Equation (9.1.3) can be written as

$$I_d = qW \frac{dV}{dx} \int_{-y}^{y} \mu_n N(y) \, dy \qquad \cdots \quad (9.1.4)$$

Assuming constant and symmetrical *n*-region doping and constant mobility, we can write equation (9.1.4) as

$$I_d = qW\,\mu_n\,N.\,2a\left(1-\frac{t}{a}\right)\frac{dV}{dx} \qquad \cdots \quad (9.1.5)$$

$$= g_0 L\left(1-\frac{t}{a}\right)\frac{dV}{dx} \qquad \cdots \quad (9.1.6)$$

where $g_0 = \left[\frac{2a\,W\,\mu_n\,qN}{L}\right]$ is the open channel conductance.

From equation (9.1.2) and (9.1.6)

$$I_d = 2\,g_0 L\,V_p\,\frac{t}{a}\left(1-\frac{t}{a}\right)\frac{d}{dx}\left(\frac{t}{a}\right) \qquad \cdots \quad (9.1.7)$$

Integrating equation (9.1.7) from $x=0$ to $x=L$ where the values of t are given by

$$t_0 = a\sqrt{\frac{V_g+V_b}{V_p}} \text{ and } t_L = a\sqrt{\frac{V_g+V_b+V_d}{V_p}} \qquad \cdots \quad (9.1.8)$$

We have with $I_0 = \frac{g_0 V_p}{3}$ $\cdots$ (9.1.9)

$$\frac{I_d}{I_0} = \left[3\frac{V_d}{V_p} - 2\left\{\left(\frac{V_g+V_b+V_d}{V_p}\right)^{3/2} - \left(\frac{V_g+V_b}{V_p}\right)^{3/2}\right\}\right] \qquad \cdots \quad (9.1.10)$$

Equation (9.1.10) applies so long there is a channel. In Fig. (9.1.2) the characteristics predicted by equation (9.1.10) are plotted. In this Figure the dotted line indicates the range of validity of equation (9.1.10) where channel pinch-off occurs due to the joint action of V_g and V_d. Therefore, on this dotted curve $(V_g+V_b+V_d)=V_p$. Hence for the case of $V_g+V_b=0$, equation (9.1.10) gives

$$\frac{I_d}{I_0} = \left[3\frac{V_d}{V_p} - 2\left(\frac{V_d}{V_p}\right)^{3/2}\right] \qquad \cdots \quad (9.1.11)$$

showing that at the pinch-off point $I_d = I_0 = \frac{g_0 V_p}{3}$. Also note, at the pinch-off points $\frac{V_g+V_b}{V_p} = \left(1-\frac{V_d}{V_p}\right)$ on each of the characteristics. Increased gate voltage, therefore, causes the pinch-off to occur at lesser source drain voltage. The characteristic $V_g+V_b=V_p$, therefore, coincides with the $\frac{V_d}{V_p}=0$ line. Equation (9.1.11) indicates that the current is obtained as the difference between an ohms law and a $\frac{3}{2}$ power law.

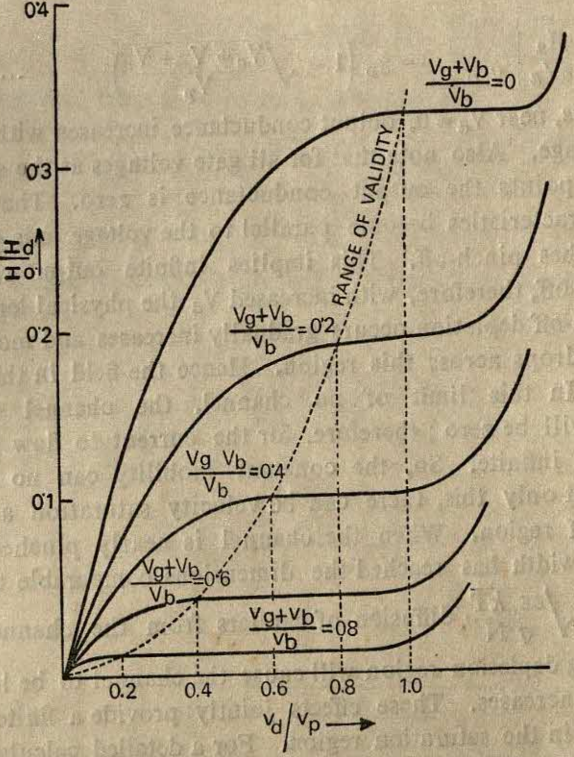


Fig. 9.1.2. Characteristics for a junction-gate field-effect transistor

The transconductance $g_m = \left.\frac{\delta I_d}{\delta V_g}\right|_{V_d \text{ Const.}}$ is obtained from equation (9.1.10) as

$$g_m = -g_0\left(\sqrt{\frac{V_g + V_b + V_d}{V_p}} - \sqrt{\frac{V_g + V_b}{V_p}}\right) \quad \cdots \quad (9.1.12)$$

at pinch-off

$$g_m = -g_0\left(1 - \sqrt{\frac{V_g + V_b}{V_p}}\right) \quad \cdots \quad (9.1.13)$$

Note that the transconductance g_m is maximum and is equal to open channel conductance for $(V_g + V_b)$ hence $V_g = 0$ since forward bias is not recommended, while it decreases with increase of the value of V_g.

As for the slope of the characteristics from equation (9.1.10), we have

$$g_d = \frac{\delta I_d}{\delta V_d}\bigg|_{V_g \text{Const.}} = g_0\left(1 - \sqrt{\frac{V_g + V_h + V_d}{V_p}}\right) \qquad \cdots \quad (9.1.14)$$

which shows, near $V_d = 0$, output conductance increases with decrease of gate voltage. Also note that for all gate voltages at the saturation (pinch-off) points the output conductance is zero. That is, the output characteristics become parallel to the voltage axis when the device reaches pinch-off. This implies infinite output resistance. After pinch-off, therefore, with increased V_d the physical length over which pinch-off depletion occurs gradually increases and most of the applied V_d drops across this region. Hence the field in this region increases. In this limit of no channel, the channel thickness $2y = 2(a-t)$ will be zero ; therefore, for the current to flow the field has to be infinite. So, the constant mobility can no more be assumed, not only this, there can be velocity saturation as well in the pinch-off region. When the channel is nearly pinched off and the channel width has reached the dimensions comparable to Debye length $L_d = \sqrt{\frac{\varepsilon\varepsilon_0 kT}{q^2 N}}$, diffusion of carriers from the channel to the neighbouring depletion region will cause the channel to be left open even if V_d increases. These effects jointly provide a finite output conductance in the saturation region. For a detailed calculation the student is referred to ref. (1). Trofimenkoff used an empirical equation for mobility and found a good fit to experimental characteristics. A similar equation may be written as

$$\mu_n = \frac{\mu_0}{1 + \frac{\mu_0 E_x}{v_s}} \qquad \cdots \quad (9.1.15)$$

where v_s is the saturation velocity of the carriers.

Substituting this in equation (9.1.5) and writing in the form of equation (9.1.6)

$$I_d = \frac{g_0 L\left(1 - \frac{t}{a}\right)\frac{dV}{dx}}{1 + \frac{\mu_0 \,|E_x|}{v_s}} \qquad \cdots \quad (9.1.16)$$

which can be written as

$$I_d\ |dx| + \frac{\mu_0\ I_d\ |dV|}{v_s} = V_p g_0 L\ \left(1 - \frac{t}{a}\right) \frac{dV}{V_p}$$

$$= g_0 V_p L\ \left(1 - \frac{t}{a}\right) d\left(\frac{t}{a}\right)^2,$$

and integrated from $x = 0$ to $x = L$ as before gives

$$I_d = \frac{g_0 V_p}{3} \frac{\left[3\frac{V_d}{V_p} = 2\left(\frac{V_g + V_b + V_d}{V_n}\right)^{3/2} - \left(\frac{V_g + V_b}{V_b}\right)^{3/2}\right]}{1 + \frac{\mu_0 V_d}{v_s L}} \quad \cdots \quad (9.1.17)$$

Equation (9.1.17) can be expressed in terms of the current of equation (9.1.10) as

$$I_d = \frac{I_d(\mu - \mu_0)}{1 + \frac{\mu_0 V_d}{v_s L}} \quad \cdots \quad (9.1.18)$$

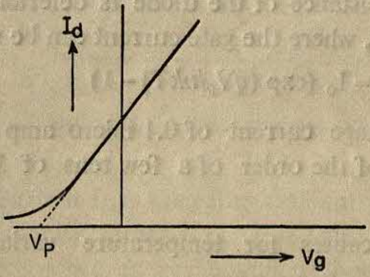


Fig. 9.1.3. Saturation region transfer characteristics JFET

In this case the drain current reaches its saturation value before pinch-off. Near pinch-off the field at the drain end becomes large; under this condition equation (9.1.16) gives

$$I_d = g_0\ \frac{L}{\mu_0}\left(1 - \frac{t_L}{a}\right) v_s \quad \cdots \quad (9.1.19)$$

which shows that the channel thickness at the drain is proportional to the current. The slope of the output characteristics due to pinch-off, therefore, can be ascribed to the variation of drift velocity in the saturation region and on the value of the resistance of the space-charge region. The resistance of the space charge region L_0 can be computed from voltage drop in the space-charge region $V = (qN\,L_0^2)/2\varepsilon\varepsilon_0$ and the current $I = qN\,v_s\,A$ as $R_{sc} = (L_0^2/2A\varepsilon\varepsilon_0\,v_s)$.

Detailed theoretical considerations and experiments indicate (1) an expression for current in the saturation region

$$I_d = I_{do}\left(1 - \frac{V_g + V_b}{V_p}\right)^2 \qquad \cdots \quad (9.1.20)$$

to hold where I_{do}, the current at $(V_g + V_b) = 0$ and V_p the pinch-off voltage, should be experimentally determined. That is, the transfer characteristic in the saturation region follows a square law as shown in Fig. (9.1.3).

Breakdown of the gate to channel diode occurs due to the joint action of V_g and V_d, because the reverse biased junction breakdown is determined by the voltage $(V_g + V_d + V_b)$. So, if V_g is small, breakdown occurs at a larger value of V_d. Since the electric field increases from source to drain and breakdown occurs at the drain end because at this end breakdown field is attained, a rough estimate of the breakdown voltage would be $V_{br} = (E_{br} . L/2)$.

The input resistance of the diode is determined by the reverse biased *p-n* junction, where the gate current can be written as

$$I_g = I_0 \left(\exp\left(qV_g/nkT\right) - 1\right) \qquad \cdots \quad (9.1.21)$$

Considering a leakage current of 0.1 micro amp at 5 volt bias, the input resistances of the order of a few tens of Megaohm is easily obtainable.

The major causes for temperature variation of the JFET parameter are

(*i*) The variation of mobility with temperature;

(*ii*) Variation of built in potential (V_b) of *p-n* junction ;

(*iii*) Variation of gate leakage current.

The result being that the drain current decreases with temperature as T^{-2} which in the usable temperature range amounts to nearly linear variation over by about 0.38% per degree celsius while the output resistance (r_{ds}) increases with rise of temperature by 0.35% per degree celsius more linearly than I_d. The result being that the amplification factor increases by about 11% for 100° celsius change for silicon material.

A figure of merit for the device can be estimated by g_m divided by the gate capacitance at pinch-off. This comes to $\frac{\mu_n q N}{\varepsilon\varepsilon_0}\left(\frac{a}{L}\right)^2$ hence high mobility doping product with short channel length and $\frac{a}{L}$ ratio improvement is necessary for improved frequency performance.

9.2 Field effect diode

If the gate of a JFET is shorted to the source then it can be used as a current limiter diode. The characteristic of the diode is the same as the JFET output characteristic with $V_g = 0$ and is as shown in Fig. (9.2.1). The important parameters of the diode are shown in this figure and these are :

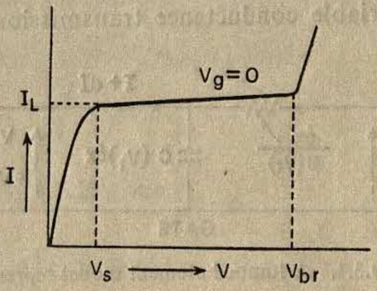


Fig. 9.2.1. Shorted source gate JFET characteristics

The limiting current I_L is the saturation current obtained due to high electric field velocity saturation. For this Ge material is more suitable than silicon since in germanium velocity saturates around 4 kv/cm where as in silicon at around 30 kv/cm. The limiting current $I_L = A\, q\, N\, v_s + I_0$ where I_0 is the diode reverse saturation current which increases with rise of temperature while v_s the saturation velocity decreases with rise of temperature. To decrease I_L the area A, doping level N or open channel width a can be decreased or L can be increased. The slope g_0 of the limiting characteristic is determined by the variation of drift velocity in saturation region and the space-charge resistance of the pinched-off region. To decrease this slope L should be increased since I_d varies inversely as L.

The saturation voltage V_s is $V_s = E_s L + I_L . R_{se}$ where R_{se} is the total series resistance due to the remaining bulk plus that due to any in the contacts while E_s is the field at the start of the saturation. To reduce the saturation voltage the channel thickness a and the doping density N should be reduced.

Lastly, the breakdown voltage V_{br} depends on the breakdown field which increases with the band gap of the semiconductor material used (*cf.* equation 4.17.5).

These diodes are available in two forms known as limiting velocity diode and FED. The saturation region width of the former is larger which causes their having less capacitance higher speed than the FED.

9.3 Small signal equivalent circuit

In this section we intend to find out the small signal equivalent circuit of the JFET. The channel region of the JFET can be regarded as a variable conductance transmission line, an elemental section of which, is shown in Fig. (9.3.1). For the channel region we can write for the series voltage drop

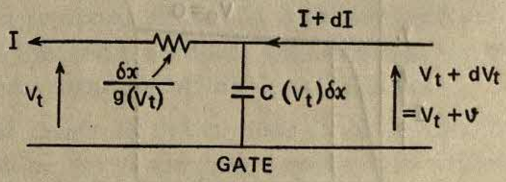


Fig. 9.3.1. A lumped element model representing a δx length of the JFET channel

$$-\delta V_t = \frac{I\delta x}{g(V_t)} \qquad \cdots \quad (9.3.1)$$

which can be written as

$$I = -g\,(V_t)\,\frac{\delta V_t}{\delta x} \qquad \cdots \quad (9.3.2)$$

For an increment of current $\delta I = i$ from (9.3.2) we can write

$$I + i = -(g + \delta g)\,\frac{\delta}{\delta x}\,(V_t + v) \qquad \cdots \quad (9.3.3)$$

With equation (9.3.2) this gives

$$i = -\left(\delta g \cdot \frac{\delta V_t}{\delta x} + g\,\frac{\delta v}{\delta x}\right)$$

or
$$i = -\left(\frac{\delta g}{\delta x} \cdot v + g\,\frac{\delta v}{\delta x}\right)$$

That is
$$i = -\frac{\delta}{\delta x}\,(gv) \qquad \cdots \quad (9.3.4)$$

On the other hand from fig. (9.3.1)

$$-\frac{\delta i}{\delta x} = j\omega c\,(V_t)\,v \qquad \cdots \quad (9.3.5)$$

From equation (9.1.6) and (9.1.1) with $z = (y/a)$

$$g = g_0 L z \qquad \cdots \quad (9.3.6)$$

With equation (9.3.4)

$$i = -g_0 L\,\frac{\delta}{\delta x}\,(zv) \qquad \cdots \quad (9.3.7)$$

Since $$\frac{\delta i}{\delta z} = \left(\frac{\delta i}{\delta x} \Big/ \frac{\delta z}{\delta x}\right) \qquad \cdots \quad (9.3.8)$$

and $z = \frac{y}{a} = (1 - \sqrt{V_t/V_p})$ *i.e.* $V_t = V_p\,(1-z)^2$ $\cdots$ (9.3.9)

gives $\frac{\delta V_t}{\delta z} = -2V_p\,(1-z)$ $\cdots$ (9.3.10)

which with equation (9.3.2) gives

$$\frac{\delta z}{\delta x} = \frac{I}{2g\,V_p\,(1-z)} \qquad \cdots \quad (9.3.11)$$

While per unit length depletion region charge is $Q = 2WtNq$

giving $C(V_t) = \frac{dQ}{dV_t} = \frac{W\,Nq\,a}{V_p\,(1-z)}$ $\cdots$ (9.3.12)

since from equation (9.1.2) we have $t = a\,\sqrt{Vt/V_p}$.

Hence we have equation (9.3.8) with (9.3.5), (9.3.11) and (9.3.12)

$$\frac{di}{dz} = -2j\omega\,\frac{WN\,qa}{I}\,gv \qquad \cdots \quad (9.3.13)$$

which with equation (9.3.6) and using $I_u = 4WN\,q\mu\,V_p$ with the expression for V_p gives

$$\frac{di}{dz} = -2j\omega\,\frac{\varepsilon\varepsilon_0 W I_u z v}{I} \qquad \cdots \quad (9.3.14)$$

Hence $$\frac{d^2 i}{dz^2} = -2j\omega\,\frac{\varepsilon\varepsilon_0 W I_u}{I}\cdot\frac{dzv}{dz}$$

which with equations (9.3.7) and (9.3.11) gives

$$\frac{d^2 i}{dz^2} + 2K^2 z\,(1-z)\,i = 0 \qquad \cdots \quad (9.3.15)$$

where $$K^2 = -j\frac{\omega\varepsilon\varepsilon_0}{2\mu\, qN}\left[\frac{I_u}{I}\right]^2 \quad \cdots \quad (9.3.16)$$

with $$I_u = 4\, Nq\mu\, WV_p \quad \cdots \quad (9.3.17)$$

and for I we have I_d of equation (9.1.11) to give (I_u/I) a non-dimensional quantity depending on external bias and pinch-off voltage.

Equations (9.3.14) and (9.3.15) describe the a.c. performance of the JFET under small signal operation. Equation (9.3.15) is a linear differential equation of second order. Hence the solution can be written as (cf. ref. 1)

$$i(K, z) = A_1\, P(K, z) + A_2\, Q(K, z) \quad \cdots \quad (9.3.18)$$

where $$P(K, z) = 1 - \frac{K^2 z^3}{3} + \frac{K^2 z^4}{6} + \frac{K^4 z^6}{45} - \frac{K^4 z^7}{42} + \cdots \quad \cdots \quad (9.3.19)$$

and $$Q(K, z) = z - \frac{K^2 z^4}{6} + \frac{K^2 z^5}{10} + \frac{K^4 z^7}{126} - \frac{K^4 z^8}{105} + \cdots \quad \cdots \quad (9.3.20)$$

By differentiating equation (9.3.18) and using (9.3.14) we have

$$v = \frac{-I/I_u}{2j\omega\varepsilon\varepsilon_0 Wz}\left[A_1\, P'(K, z) + A_2\, Q'(K, z)\right] \quad \cdots \quad (9.3.21)$$

Using equation (9.3.18) and circuit conventions of current, we can write for the terminal currents with appropriate subscript for boundary conditions, with z and currents at source and drain ends as,

$$i_s = -A_1\, P(K, z_s) - A_2\, Q(K, z_s) = -A_1 P_s - A_2 Q_s \quad \cdots \quad (9.3.22)$$

$$i_d = -A_1\, P(K, z_d) + A_2\, Q(K, z_d) = A_1 P_d + A_2 Q_d \quad \cdots \quad (9.3.22a)$$

hence $$i_g = A_1(P_s - P_d) + A_2(Q_s - Q_d) \quad \cdots \quad (9.3.23)$$

where $P_s = P(K, z_s)$ and $Q_s = Q(K, z_s)$ and so on.

Again from equation (9.3.21) and fig. (9.3.1)

$$v_{gs} = \frac{-I/I_u}{2j\omega\varepsilon\varepsilon_0 Wz_s}(A_1 P_s' + A_2 Q_s') \quad \cdots \quad (9.3.24)$$

$$v_{dg} = -v_{gd} = \frac{(I/I_u)}{2j\omega\varepsilon\varepsilon_0 Wz_d}(A_1 P_d' + A_2 Q_d') \quad \cdots \quad (9.3.25)$$

where dash refers to differentiation of the series prior to putting boundary values.

While $v_{gs} + v_{sd} + v_{dg} = 0$ hence,

$$v_{sd} = \frac{I/I_u}{2j\omega\varepsilon\varepsilon_0 W}\left(\frac{A_1}{z_s}P_s' - \frac{A_1}{z_d}P_d' + \frac{A_2}{z_s}Q_s' - \frac{A_2}{z_s}Q_d'\right) \quad \cdots \quad (9.3.26)$$

To find the intrinsic parameters for common source equivalent circuit, one then evaluates the constants corresponding to the appropriate boundary conditions. For example, for y_{11i} and y_{21i} we know

$$y_{11i}=\left(\frac{i_g}{v_{sg}}\right)_{v_{sd}=0} \quad \text{and} \quad y_{21i}=\left(\frac{i_d}{v_{sg}}\right)_{v_{sd}=0} \quad \cdots \quad (9.3.27)$$

Hence equating (9.3.26) to zero one finds the relationship between A_1 and A_2 which, on using with equations (9.3.23), (9.3.24) and (9.3.22), one obtains, from equation (9.3.27)

$$y_{11i}=\frac{2j\omega\varepsilon\varepsilon_0 WI_u}{I}\left[\frac{(P_s-P_d)(Q_d'z_s-Q_s'z_d)+(Q_s-Q_d)(P_s'z_d-P_d'z_s)}{P_s'Q_d'-Q_s'P_d'}\right] \quad \cdots \quad (9.3.28)$$

$$\text{and} \quad y_{21i}=\frac{2j\omega\varepsilon\varepsilon_0 W I_u}{I}\left[\frac{P_d(Q_d'z_s-Q_s'z_d)+(P_s'z_d-P_d'z_s)Q_s}{P_s'Q_d'-Q_s'P_d'}\right] \quad (9.3.29)$$

Similar approach can be used to find y_{12i} and y_{22i}.

After finding the intrinsic y parameters of the black box if one aims at finding an equivalent circuit of the form, as shown in fig. (9.3.2), then the open circuit and short circuit admittances can be equated and manipulated to give

$$\left.\begin{aligned} y_{2i}&=-y_{12} \\ y_{oi}&=y_{22}+y_{12} \\ y_{mi}&=y_{21i}-y_{12} \\ y_{1i}&=y_{11i}+y_{12} \end{aligned}\right\} \quad \cdots \quad (9.3.30)$$

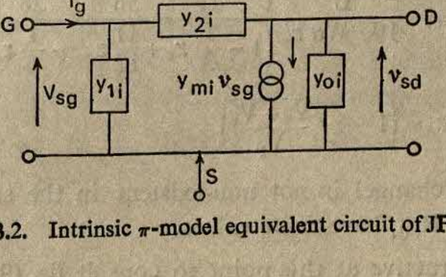


Fig. 9.3.2. Intrinsic π-model equivalent circuit of JFET.

Using equations (9.3.30) with the values of intrinsic admittances found by equations (9.3.28), (9.3.29) and the like, retaining terms upto ω^2 after a lengthy piece of algebra, one obtains an equivalent circuit of the form as shown in fig. (9.3.3) for the intrinsic transistor ;

where, assuming the output conductance as zero in the saturation region, the values of the parameters are as given by Cobbold.(1)

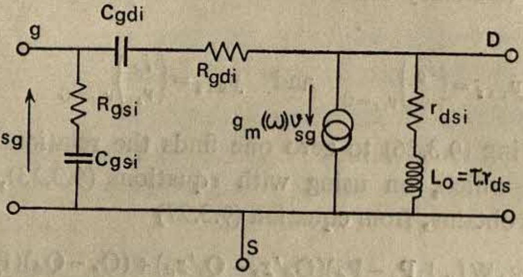


Fig. 9.3.3. Intrinsic equivalent circuit of JFET

$$g_m(\omega) = \frac{2\sigma W a\, z_s}{L} / (1 + j\omega\, T_o)$$

$$T_o = \frac{4\varepsilon\varepsilon_0 L^2}{15\sigma a^2} \frac{(1 - 5/4 z_s + 5/14 z_s^2)}{z_s(1 - 2/3\, z_s)^3}$$

$$C_{gsi} = \frac{4W\varepsilon\varepsilon_0 L}{3a} \frac{(1 - z_s/2)}{(1 - \frac{2}{3} z_s)^2} \qquad \cdots \quad (9.3.31)$$

$$C_{gdi} = 0$$

$$r_{dsi} = \infty$$

$$R_{gdi} = \infty$$

$$L_0 = \infty$$

$$R_{gsi} = \frac{L}{10\sigma\, W a\, z_s} \frac{\left(1 - \frac{37}{21} z_s + \frac{55}{56} z_s^2 - \frac{5}{28} z_s^3\right)}{\left(1 - \frac{5}{3} z_s + \frac{11}{12} z_s^2 - \frac{1}{6} z_s^3\right)}$$

with $$z_s = \left(1 - \sqrt{\frac{V_g + V_b}{V_p}}\right)$$

whereas since channel is not non-existent in the saturation region, therefore, r_{dsi}, R_{gdi} have finite values.

It is instructive at this point to consult fig. (9.3.4) to visualise the origin of the components of fig. (9.3.3) as also of the extrinsic elements.

The interelectrode capacitances should be taken into consideration, over and above the inclusion of the extrinsic elements shown in fig. (9.3.4) ; an important one of which is the effect of R_s is to modify the intrinsic g_m to $[g_m/(1+g_m R_s)]$. For frequencies above 200 MHz the lead inductances should be also taken into account.

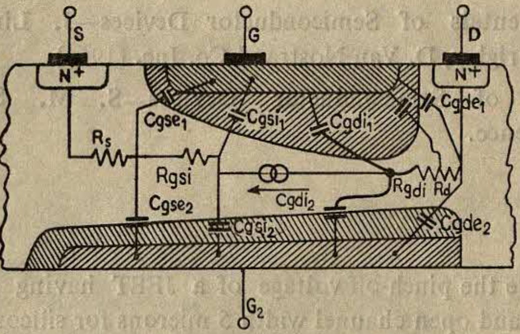


Fig. 9.3.4. Shows the origin of the components of fig. 9.3.3.

It would be seen from equation (9.3.31) that the intrinsic transistor g_m cut-off radian frequency $(1/T_0)$ is proportional to (a^2/L^2). That is to improve the frequency performance (a/L) has to be large, *i.e.*, short channel length with increased pinch-off voltage since $V_p=(qN\,a^2/2\varepsilon\varepsilon_0)$. While to improve low frequency value of g_m, $W(a/L)$ and σ can be increased but increase of σ would increase V_p, improve the input time constant and would improve low frequency g_m though increase of W does not affect the input time constant. Considering all these one looks for high mobility material allowing to use a lower doping thus preserving low V_p. But unless one goes to have channel lengths less then microns the performance of JFETs cannot be substantially improved. With silicon typical values, for medium and high frequency transistors with channel lengths down to one micron including the extrinsic parameters are as :

g_m in m-mho	C_{gs} in *pf*	R_{gs} in ohms	r_{ds} ohms	R_{gd} ohms	f_T GHz	$C_{o\,ex}$ *pf*
2	10	50	80K	40	0.2	0.5
5	3	20	50K	6	1.6	0.3

To improve the frequency performance therefore, submicron range channel lengths has to be used with materials like GaAs which have higher mobility.

References :

1. Theory and Applications of Field Effect Transistors—Richard S. C. Cobbold, Wiley-Interscience (1970)
2. Fundamentals of Semiconductor Devices—J. Lindmayer and C. Y. Wrigley, D. Van Nostrand Co. Inc. (1965)
3. Physics of Semiconductor Devices—S. M. Sze, Wiley-Interscience.

Problems :

1. Calculate the pinch-off voltage of a JFET having N-doping of 10^{18}/cc and open channel width 5 microns for silicon material.
2. Calculate the g_m of a JFET having a gate voltage of 2 volts, built in voltage 0·5 volts and a pinch-off voltage of 5 volts, assume open channel conductance 2 mA/volts.
3. Find the output conductance of the JFET of Problem No. 2 above for an applied gate voltage 2 volts and drain source voltage of 1 volt.

Chapter 10

SEMICONDUCTOR TECHNOLOGY

In this chapter we get a cursory glance of the semiconductor technological methods and process steps for fabrication of devices that will be useful for the understanding of the devices. The basic technological methods that are useful for fabrication of semiconductor devices are : (1) Diffusion, (2) Oxidation, (3) Epitaxy, (4) Photolithography and (5) Metallisation. We discuss here each of these processes separately and later show how in the fabrication of devices we use them. For oxidation or diffusion one normally takes a silicon slice having resistivity between 0·5—50 ohm-cm as needed for the purpose.

These lapped and polished wafers are rinsed in the deionised (DI) water of resistivity 10—20 Mohm-cm. These are then cleaned ultrasonically with acetone, trichloroethylene, again acetone and finally with absolute alcohol. Then these are dipped in H.F. solution (1 : 10) for about a minute thus ensuring the removal of thin oxide layers, if any, from the surface. After this the slices are washed in DI water. Next these are dipped in chromic acid solution at 90°C (±0·5°C) for about 15 minutes, rinsed in DI water and dried in N_2 gas stream. It is to be remembered that after a cleaning process only, the oxidation, diffusion or epitaxy follows.

10.1 Diffusion :

Diffusion is a process step that is used for doping a semiconductor. In this a semiconductor sample in the form of a slice is firstly lapped, mirror polished and, subsequently, cleaned before introducing within the quartz process tube in a furnace held at a higher temperature in the range 950°—1250°C. The furnace temperature is maintained within a close tolerance preferably within ±0·1°C. Through this furnace tube inert pure gas like N_2 or Agon used as carrier gas of the impurity is allowed to flow since prior to introducing the sample. The dopant material can be either in a solid slice form such as Boron nitride or it can be in a liquid form as BBr_3 or $POCl_3$. In the first case, *i.e.*, if the dopant material is solid, then it is placed in the process tube of the furnace adjacent to the slice on a quartz boat, while the carrier gas enters the process tube and flows at a certain moderate rate. While for liquid dopant case, the inert gas is bubbled through the liquid held in a test tube maintained at a constant low temperature and led to enter the furnace. The sample is kept in the furnace over a length of time that is required for the necessary depth of diffusion of the dopant atoms.

In diffusion, particles flow from high concentration to low concentration regions. The particle flux is given by $J = -D \text{ grad } N$ where we have neglected any field flow since at the high temperature diffusion, we are discussing, the internal electric field can be neglected. Combined with the continuity equation we have

$$\frac{\delta N}{\delta t} = D \nabla^2 N \quad \cdots \quad \cdots \quad (10.1.1)$$

Ignoring the three dimensional effects, we consider only unidirectional flow. This implies that limitation of boundaries on the other two directions would be there. If the impurity is delivered to the surface in vapour form diluted with a carrier gas or a heavily doped polycrystalline silicon deposited on the silicon sample surface, then for the solution of the one dimensional form of the above equation we can use constant concentration N_0 at the surface as the boundary value. If we assume that the slice was earlier uniformly doped to level N_A, it is semi-infinite in extension and that the N_A atoms diffuse very little compared to the new ones, we have for the solution of the one dimensional diffusion case

$$N = N_0 \text{ erfc } (x/2\sqrt{Dt}) - N_A \quad \cdots \quad \cdots \quad (10.1.2)$$

which is seen to be a function of time, also of depth.

where $\text{erfc } (x) = 1 - \text{erf } (x)$

and $\text{erf } (x) = \frac{2}{\pi} \int_0^x \exp(-z^2)\, dz$

hence $\text{erf } (o) = 0, \text{erf } (\infty) = 1$

the complementary error function profile is shown in fig. (10. 1. 1). From the steep fall within the sample it is apparent that this profile would be usable where shallow high to low doping transition is necessary such as for emitter of a BJT or a first deposition waiting for a drive in diffusion.

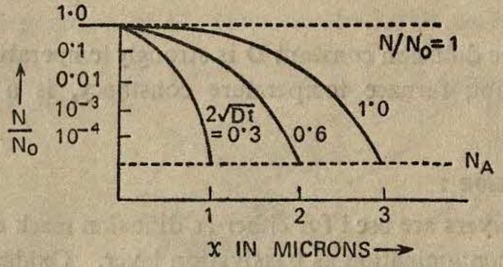


Fig. 10.1.1. Constant concentration diffusion profiles ; $2\sqrt{Dt}$ is a parameter.

The junction is formed at a depth where the total concentration given by equation (10.1.2) is zero, *i.e.*, at

$$x_j = 2\sqrt{Dt}\,\mathrm{erfc}^{-1}\left(\frac{N_A}{N_o}\right) \quad \cdots \quad \cdots \quad (10.1.3)$$

If, on the other hand, in the limited source diffusion, that is, if predeposited number of impurity atoms ΣN_o per unit area on the surface of the sample is fixed and the layer thickness T is very much thinner than the wanted depth of penetration, *i.e.*, larger times are allowed for diffusion or in other words $x < (Dt/T)$ and $t > \frac{T^2}{D}$ in that case the result of drive-in can be written as

$$N(x.\ t) = \frac{\Sigma N_o}{\sqrt{\pi D t}} \exp\left(-\frac{x^2}{4Dt}\right) \quad \cdots \quad (10.1.4)$$

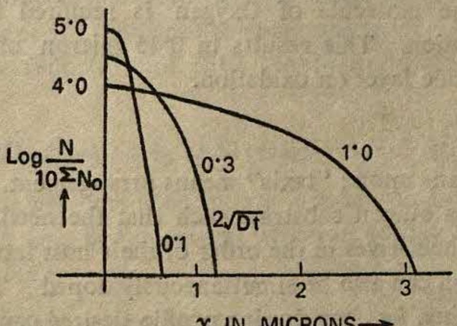


Fig. 10.1.2. Limited source diffusion profiles with $2\sqrt{Dt}$ as a parameter.

which is known as the Gaussian distribution and is as shown in fig. (10. 1.2). As can be seen from the nature of boundry conditions and its conditions of validity that this diffusion is usuable where deep diffusion is necessary, for example, isolation diffusion in I.C. fabrication.

Since the diffusion constant D is strongly temperature dependent so the diffusion furnace temperature constancy, is a must to be achieved.

10.2 Oxidation :

Oxide layers are used for either as diffusion mask or for protection against contamination as passivation layer. Oxidation is of two types, dry oxidation and wet oxidation. Dry oxidation is a slow process while wet oxidation is much faster comparatively with sacrifice of quality. For wet oxidation the wafer is placed in the furnace tube held at a constant temperature between 900°C—1200°C and oxygen as mixed with steam is passed over the slice. In dry oxidation only oxygen is used and no steam. The oxygen diffuses through the oxide formed and reaction proceeds as

$$Si + O_2 \rightarrow SiO_2$$

while with steam

$$Si + 2H_2O \rightarrow SiO_2 + 2H_2$$

$$2H_2 + O_2 \rightarrow H_2O$$

In these the oxygen atoms enter the lattice and a portion of the surface of silicon is consumed to form SiO_2 so this growth need not be confused with deposition. It is to be remembered that previously diffused impurities obviously will diffuse further while oxidation is performed. One molecule of Oxygen is required for each SiO_2 molecule formation. This results in 0·45 micron of Si giving on micron thick oxide layer on oxidation.

10.3 Epitaxy :

"Epi" means upon ; "taxis" means arrangement. In this silicon is grown over a silicon substrate such that the newly born silicon atoms arrange themselves in the order of their host lattice. The layer thus being grown can also be simultaneously doped. Thus no doping to uniform doping, *i.e.*, any doping profile desired can be achieved by this process.

The process is carried out in an epitaxial reactor, as shown in fig. (10.3.1), which consists of a quartz tube containing a silicon carbide coated graphite boat heated by R.F. power from outside such that the boat is heated only while the tube remains cold, thus avoiding possible deposition of Si atoms on the inside walls of the tube and subsequent contamination with the wafers on the boat. A mixture of silicon tetrachloride and Hydrogen gas is passed through the tube. The slices are kept at a constant temperature around 1200°C. In the overall reaction reduction of $SiCl_4$ to Si and formation of HCl occur.

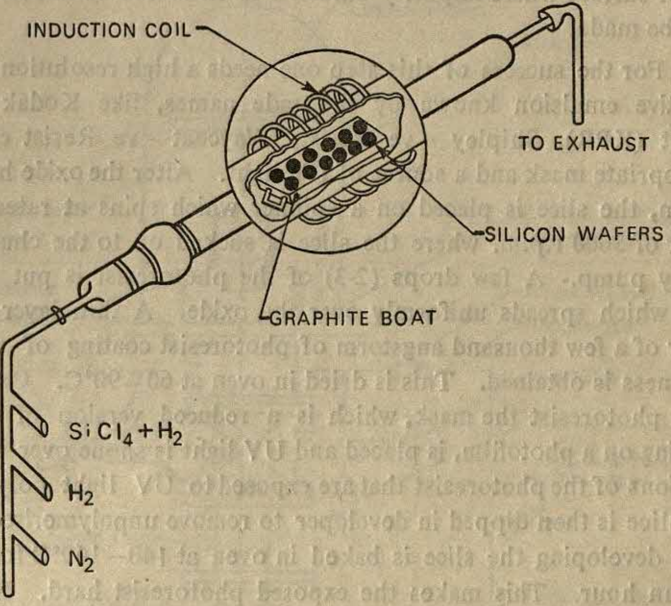


Fig. 10.3.1. Growth of epitaxial films

$$SiCl_4\ (gas) + 2H_2 \rightleftharpoons Si\ (solid) + 4\ HCl\ (gas)$$

Temperature and gas composition are critical because the reaction occurs bothways *i.e.*, in the reverse reaction freshly deposited Si is etched away by HCl.

Epitaxial layer grown on silicon substrate is known as Epi-slices. These are useful for IC fabrication.

10.4 Photolithography :

This is one of the key processes that is repeatedly to be used for device fabrication for discrete devices as also for IC's. This is mainly used for opening windows on grown oxide layers and oxides of these windows are etched away to expose the underlying semiconductor surface where impurity diffusion or oxidation or metallisation may be made.

For the success of this step one needs a high resolution photosensitive emulsion known by the trade names, like Kodak Photo Resist (KPR), Shipley +ve Resist, Waycoat −ve Resist etc., the appropriate mask and a source of UV light. After the oxide has been grown, the slice is placed on a spinner which spins at rates of the order of 3000 r.p.m. where the slice is sucked on to the chuck by a rotary pump. A few drops (2-3) of the photoresist is put on the slice which spreads uniformly over the oxide. A thin layer of the order of a few thousand angstorm of photoresist coating of uniform thickness is obtained. This is dried in oven at 60−90°C. On to this dried photoresist the mask, which is a reduced version of a large drawing on a photofilm, is placed and UV light is shone over it. The positions of the photoresist that are exposed to UV light polymerise. The slice is then dipped in developer to remove unpolymerised resist. After developing the slice is baked in oven at 140—160°C for about half an hour. This makes the exposed photoresist hard. Now the baked wafer is dipped in 'buffered hydrofloric acid' etch and selective etching is made to expose the underlying semiconductor. Then it is rinsed in deionised (DI) water followed by washing in stripper solution that strips off the photoresist. The wafers are then cleaned and finally rinsed in DI water and dried in N_2 atmosphere.

Thus we see that to expose a portion of semiconductor, we need a mask which allows light not to pass through these portions. Such a photoresist is called a negative photoresist. Positive photoresist is also available in the market which exposes portions exposed to UV light because these portions dissolves on developing.

10.5 Metallisation :

Metallisation means metal coating on to the different regions. This coating of metal is performed by vacuum deposition or by sputtering, vacuum evaporation is performed at a pressure of 10^{-6}—

10^{-7} mm of Hg. The metal, frequently aluminium, to be evaporated is held in a boron nitride or tantalum diboride crucible inside a glass or a metal chamber which is evacuated continuously from beneath, as shown in fig. (10.5.1), by rotary and diffusion pumps in series. Heating of the evaporant is done by R.F. induction or by electron

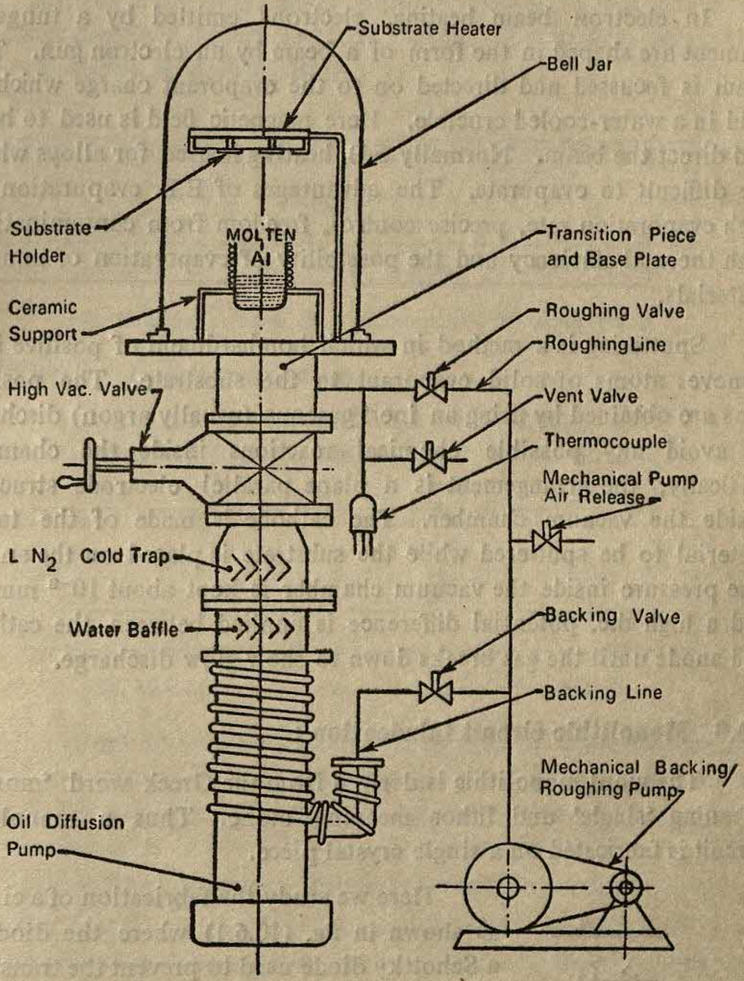


Fig. 10.5.1. High vacuum system

beam to avoid contamination. Resistance heater should be avoided for production of IC's. The crucible is supported on ceramic pillars and surround by R.F. coil tubing which carries water through it.

This is to keep the surrounding cool ; othewise, outgassing from condensed deposit and various materials in the chamber will increase the pressure above 10^{-5} mm of Hg when evaporated material will scatter away from the desired target and contamination will also result.

In electron beam heating, electrons emitted by a tungsten filament are shaped in the form of a beam by an electron gun. The beam is focussed and directed on to the evaporant charge which is held in a water-cooled crucible. Here magnetic field is used to bend and direct the beam. Normally E.B. heating is used for alloys which are difficult to evaporate. The advantages of E.B. evaporation are high evaporation rate, precise control, freedom from contamination, high thermal efficiency and the possibility of evaporation of unusual materials.

Sputtering is a method in which bombardment of positive ions removes atoms of solid evaporant to the substrate. The positive ions are obtained by using an inert gaseous (usually argon) discharge to avoid any possible chemical reactions inside the chamber. Basically, the arrangement is a plane parallel electrode structure inside the vacuum chamber. The cathode is made of the target material to be sputtered while the substrate is placed on the anode. The pressure inside the vacuum chamber is kept about 10^{-2} mm Hg and a high d.c. potential difference is applied between the cathode and anode until the gas breaks down to show glow discharge.

10.6 Monolithic circuit fabrication :

The word monolithic is derived from the Greek word 'monos', meaning 'single' and lithos meaning stone. Thus a monolithic circuit is fabricated on a single crystal piece.

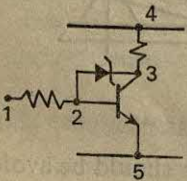


Fig. 10.6.1

Here we study the fabrication of a circuit as shown in fig. (10.6.1) where the diode is a Schottky diode used to prevent the transistor from entering into deep saturation thus enabling a reduction in the saturation delay in switching. To fabricate this circuit on a monolithic chip the technological steps are as described on next page (p. 251).

Step 1. Oxide Growth* :

An (111) epitaxial silicon slice having for P-type substrate 6—8 mil thickness and 10—50 ohm-cm resistivity is chosen with 6—8 micron N-epi-layer thickness of resistivity 10—15 ohm-cm for the purpose. This slice is cleaned as described before and then a thin (~0.5 micron) oxide SiO$_2$ layer is grown over it by wet oxidation process as described in article (10.2). After this process step, the slice is as shown in fig. 10.6.2(a).

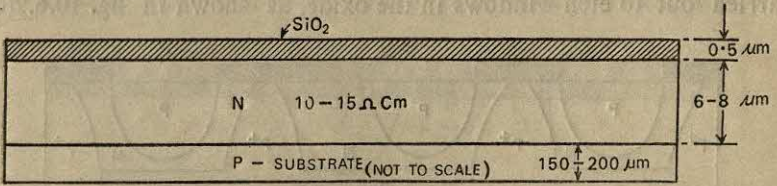


Fig. 10.6.2(a). Shows the dimensions and the oxide on top.

Step 2. Isolation diffusion :

By using a suitable mask with a photolithographic process step, the oxide is removed from certain places as shown in fig. 10.6.2(b)

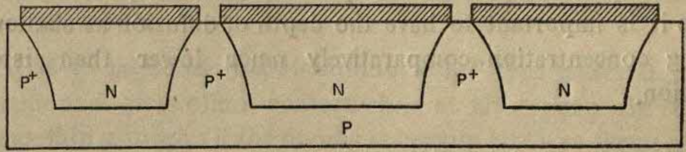


Fig. 10.6.2(b). Shows the isolated island P+ diffusion.

and after cleaning as described in article (10.4) P+-diffusion is made through the cut-windows so that isolated N-islands are obtained as shown in the diagram. The P-type substrate will be held at the most negative potential so that the N-type islands will behave in circuit as isolated by the high reverse biased resistance to substrate.

The doping in isolation diffusion is high ($P^+ \sim 0.5 \times 10^{20}$) so that the depletion regions extend into the lighter N-doped regions thus avoiding possible shorts of N-islands. It is to be noted that the

---

[*] To reduce the collector series resistance for $V_{ee(sat)}$ lowering or as required for H.F. transistors 'buried layer' diffusions are made. These are so located that after the epi-layer growth these regions fall beneath the collector regions. For IC's in particular such epi-slices are used.

N-islands thus have distributed parasitic capacitances to substrate which affect the speed of operation in digital circuits and slew rate in linear circuits. The P-diffusion concentration is chosen to limit this capacitance to a few picofards only.

Step 3. Base diffusion :

After isolation diffusion a layer of oxide is formed over the entire surface and with suitable mask a photolithographic step is carried out to etch windows in the oxide, as shown in fig. 10.6.2(c).

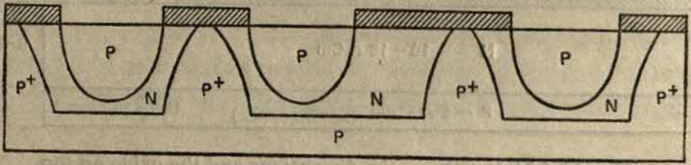


Fig. 10.6.2(c). Shows the sample after P base diffusion.

Then after usual cleaning the P-type base diffusion is made. In this step the base region, the resistor regions, the anode of diodes and junction capacitors (if to be incorporated in circuits) are catered for. Hence it is important to have the depth of diffusion as shallow the doping concentration comparatively much lower than isolation diffusion.

Step 4. Emitter diffusion :

A new layer of oxide is again formed on the entire surface. Again with suitable mask a photolithographic step is carried out to etch windows for emitter region, collector metallisation regions or diode cathode regions and capacitors (if any). Through these windows shallow heavy N-type phosphorous diffusion is made after completion of which it is as shown in fig. 10.6.2(d).

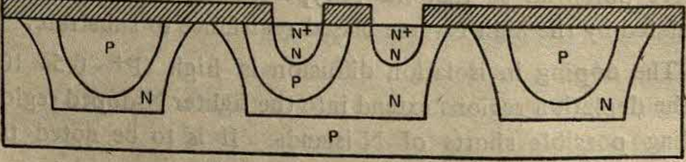


Fig. 10.6.2(d). The sample after 'N'-diffusion prior to further oxide growth.

Step 5. Metallisation :

Another layer of oxide is again grown on the entire surface and with a suitable mask a photolithographic step is carried out to etch windows for metallised coatings of the component electrode forming regions. Aluminium is coated by vacuum evaporation to the entire surface which is then etched away in another photolithographic step with a suitable mask at selected regions to give pattern like fig. 10.6.2(e). This is how the circuit mentioned has been realised.

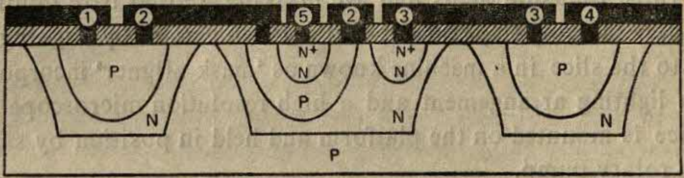


Fig. 10.6.2(e). Shows the circuit after metallisation.

The slices are then sintered in N_2 atmosphere at about 580°C for 10—15 minutes. The aluminium forms eutectic mixture with local silicon underneath.

It is to be noted that aluminium is a P-type material ; hence on P-regions it gives ohmic contact while at N^+ regions the barrier becomes thin enough (if the doping is heavily N^{++}) so forms also an ohmic contact where the carriers tunnel through the thin barrier which is of the order of atomic dimensions. Basically, it becomes a metal to degenerate semiconductor to N-type semiconductor transition. Also note that the positive electrode for Schottky is Al which connects the N-type collector region to form the Schottky barrier, *i.e.*, when N-side that is collector is more negative than the base P, the forward current for Schottky flows through the metal Al instead of through semiconductor. Individual circuits or devices are then tested and the underrated ones are marked with special ink spots.

Step 6. Die bonding :

The slices are then put in a scribing machine, the device boundaries are marked with a fine diamond tool and the individual device chips are separated out. Each chip is mounted on a suitable header and to the posts of the header 1 mil Al or gold wires are stitch bonded from the required metallised terminal pads.

It is to be noted that in Step 1 we used a silicon slice with (111) plane as the flat surface. This is because the (111) plane has got the highest mobility of minority carriers. In bipolar technology this is the preferred orientation of crystal plane.

Also note that the technological steps described above is the standard bipolar technology and that this requires four oxidation steps, three diffusion steps and five masking steps as bare minimum. That these masking steps do not foul is taken care of by aligning each mask to the slice in a machine known as 'mask aligner' incorporated with a lighting arrangement and a high resolution microscope while the slice is mounted on the platform and held in position by suction from a rotary pump.

References :

1. Integrated Circuit Engineering—Glaser, Subak-Sharpe, Addison Wesley Publishing Co.
2. Physics of Semiconductor Devices—S. M. Sze, Wiley-Interscience.

Chapter 11

MOS AND MOSFET

In this chapter Metal to Oxide to Semiconductor interface and Field effect transistor based on carrier flow on this type of semiconductor interface will be discussed. In this structure the oxide of thickness greater than 100°A, normally of the order of 1000°A acts as an insulator grown on the semicoductor ; usually as SiO_2 on silicon and on top of this oxide a metal electrode called gate is deposited. This metal gate and an ohmic contact on the semiconductor forms a capacitor element. The application of a d.c. bias on these electrodes causes a field to appear at the oxide semiconductor interface region. The effect of this field is the subject of study of this chapter.

11.1 MOS Structure :

If we have started with a *p*-type semiconductor and the metal is made negative with respect to the semiconductor, then the interface field is directed towards the oxide from the semiconductor pulling in electrons within the bulk from the interface regions and pushing holes to the interface as shown in fig. 11.1.1(a). If the applied voltage is appreciable we would expect the effect to be considerable and there would be appreciable increase of holes (the majority carriers) at the

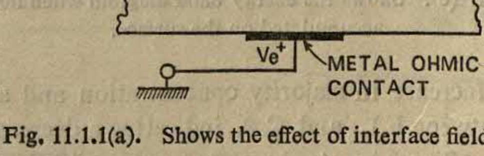


Fig. 11.1.1(a). Shows the effect of interface field.

surface with a reduction in the number of minority carriers. In this state we speak of the semiconductor surface as an accumulated one, when due to accumulation of majority carriers the interface region conductivity increases than that of the bulk region as shown in fig. 11.1.1(b). Fig. 11.1.1(c) depicts the expected energy band diagram

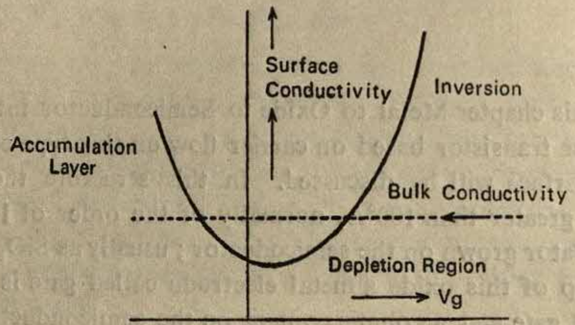


Fig. 11.1.1(b). Surface conductivity *vrs* gate voltage.

during this situation from the bulk to the surface of the semiconductor. Deep within the bulk semiconductor the carrier concentration hence the distance between the F.L. and the V.B. has not changed. As the surface is approached on the left this distance decreases

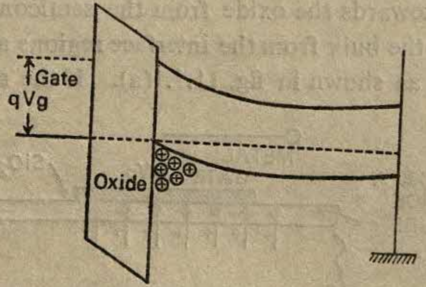


Fig. 11.1.1(c). Shows the energy band diagram when holes are accumulated on the surface.

indicating an increase in majority concentration and an increase in the distance between F.L. and C.B. indicating a decrease of minority concentration. Since in steady state an equilibrium is reached, therefore, the F.L. is shown as flat between the metal through semiconductor to ground while the applied voltage difference between ground to gate metal has got distributed partly in the oxide layer and partly in the semiconductor through the interface region.

If the gate voltage is increased in the positive direction from the accumulated state, the field strength decreases, the surface conductivity decreases and at a certain voltage it becomes minimum. The value of this voltage obviously depends on the metal work function, semiconductor electron affinity and doping level apart from the state of the surface, which may have carrier capturing centres known as surface states, since it is a surface and is different from the bulk semiconductor in respect of microscopic electrical forces of the atom cores and others. The minimum of conductivity $q\left(\mu_p p + \mu_n \frac{n_i^2}{p}\right)$ will be attained when the surface hole concentration reaches $p = n_i \sqrt{\mu_n/\mu_p}$ and electron concentration $n = n_i \sqrt{\mu_p/\mu_n}$ as it happens normally in the bulk semiconductor apart from the differences occurring due to different values of the mobilities on the surface from the bulk region.

If the gate voltage is made more positive, more holes move out of the interface region and the region becomes depleted of mobile holes while minority carrier electron concentration increases. With increasing gate voltage, therefore, the depletion region width increases while minority carrier concentration at the surface increases ultimately becoming much more than the majority carrier concentration at the surface, and the surface conductivity also increases with voltage. When the surface minority carrier concentration becomes equal to the bulk majority carrier concentration, the surface is said to be inverted, *i.e.*, the surface conductivity is primarily determined by minority carriers instead of majority carriers. In this state the

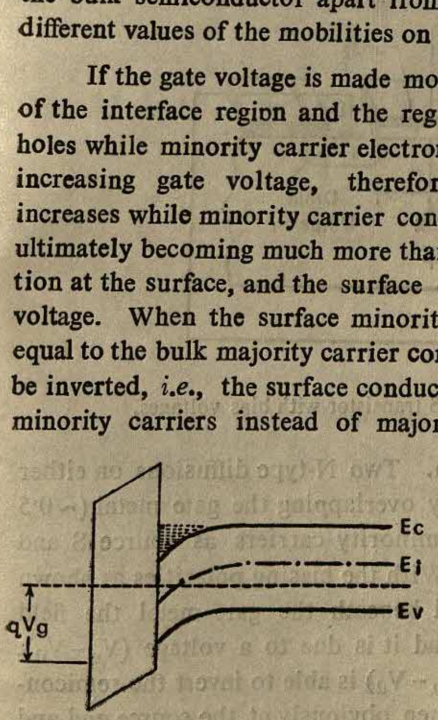


Fig. 11.1.1(d). Energy band diagram when SiO_2-Si interface region is inverted.

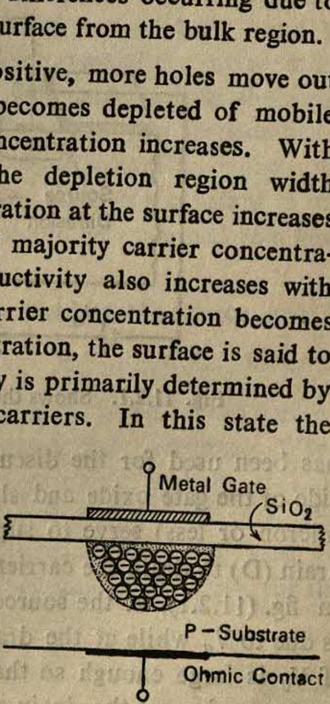


Fig. 11.1.1(e). Shows the inverted mobile charge at the SiO_2-Sem. interface and depleted fixed charge beneath this.

energy band diagram will look like that as shown in fig. 11.1.1(d) and surface charge is shown in fig. 11.1.1(e) where the mobile minority carriers are mostly available at the semiconductor surface while immobile acceptor sites extend to depths of the order of a micron or so. Similar phenomena will also happen if the substrate was N-type while the bias polarities and carrier types were oposite to that used in the above discussion.

11.2. MOS Transistor :

The MOS transistor or MOST utilises the principle of carrier depletion, inversion or accumulation creating ability of the gate voltage on the surface of a semiconductor at the oxide semiconductor interface, the gate being used as the control electrode. A typical transistor structure is shown in fig. (11.2.1) where a P-type substrate has been used for the discussion. Two N-type diffusions on either side of the gate oxide and slightly overlapping the gate metal (~0·5 micron or less) serve to supply minority carriers as source S and drain (D) to drain the carriers. With the biasing polarities as shown in fig. (11.2.1), at the source end beneath the gate metal the field is due to V_g while at the drain end it is due to a voltage $(V_g - V_d)$. If V_g is large enough so that $(V_g - V_d)$ is able to invert the semiconductor surface at the drain end then obviously at the source end and throughout the gate length on the surface inversion exists. In that case, the source and the drain regions get connected by the inverted layer of mobile charges (electrons) and current flows between source

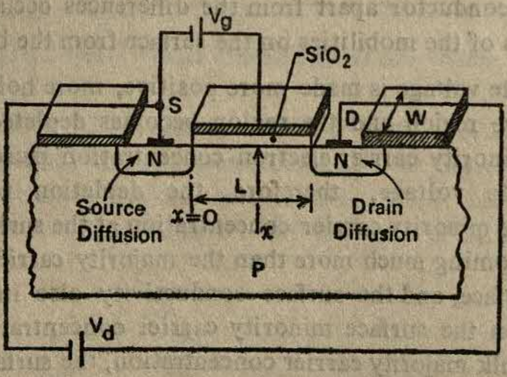


Fig. 11.2.1. Shows the MOS transistor with bias voltages.

and drain. Variation of V_g or V_d or both causes the surface electron concentration ; hence conductivity of the surface layer to change that changes the current. This is the basic principle of operation of the MOS transistor.

To understand the nature of the current voltage characteristics quantitatively, the field E, that exists on the semiconductor surface can be written in terms of the surface charge per unit area from Gauss' law as

$$\varepsilon E = Q_{sur} \qquad \cdots \quad (11.2.1)$$

If the insulator thickness is t then it can be written as

$$E = -\frac{V_g - V_x - V_T}{t} \qquad (11.2.2)$$

where V_x is the running voltage on the surface at a distance x measured with respect to source end at $x = 0$ as shown in fig. (11. 2. 1) and V_T is a voltage which takes into account of the effect of the surface state, the surface being different from the bulk and because the work function difference of the metal, insulator and the semiconductor should be taken into account. It is to be noted that throughout the whole length of the surface, V_T has been taken as a constant in our discussion, the composition of which would be examined in a later article.

From equation (11.2.2) the surface field at the source end is $\left(\frac{V_g - V_T}{t}\right)$ while at the drain end it is $(V_g - V_T - V_d)/t$, *i.e.*, lower at the drain end. Hence for a certain V_g, as V_d increases the field on the surface throughout as also at the drain end decreases. When $V_g - V_T = V_d$ the field at the drain end becomes zero, and by (11.2.1) the channel charge at the drain end becomes zero. That is, the channel becomes non-existent or saturates. Hence so long as $V_g - V_T > V_d$ the channel exists and in that case we can write for the channel current flowing from drain to source as :

$$I_d = \mu \frac{dV_x}{dx} W\, Q_{sur} \qquad \cdots \quad (11.2.3)$$

where W is the width as shown in fig. (11.2.1) and μ is the carrier mobility on the surface. Substituting for Q_{sur} from equation (11.2.1) using (11.2.2) and rearranging, we have

$$I_d \cdot dx = \frac{\mu W \varepsilon \varepsilon_0}{t} (V_g - V_T - V_x)\, dV_x \quad \cdots \quad (11.2.3a)$$

which on integration between the source end at $x = 0$ where $V_x = 0$ to the drain end at $x = L$ where $V_x = V_d$ gives

$$I_d = \frac{\mu W \varepsilon \varepsilon_0}{tL} \left[V_d (V_g - V_T) - \frac{V_d^2}{2} \right] \quad \cdots \quad (11.2.4)$$

Equation (11. 2. 4) gives the drain current so long $V_g - V_T \gg V_d$ and the channel exists. The channel saturates when $V_g - V_T = V_d$ when the current saturates and becomes

$$I_{d\,sat} = \frac{\mu W \varepsilon \varepsilon_0 V_d^2}{2tL} = \frac{\beta}{2} V_d^2 = \frac{\beta}{2} (V_g - V_T)^2 \quad \cdots \quad (11.2.5)$$

$$\text{where} \quad \beta = \frac{\mu W \varepsilon \varepsilon_0}{Lt} \quad \cdots \quad (11.2.6)$$

Equation (11.2.5) shows that the transfer characteristics follows a square law which leads to the use of the device as a squarer. Note that the term saturation in MOS refers to the absence of channel and is used with a meaning different from the word saturation of a bipolar transistor. The current voltage characteristic is shown in fig. (11.2.2) in which the saturation boundary is shown with a chain dotted curve. Note that $V_g = V_d$ curve implies always in saturation in this figure.

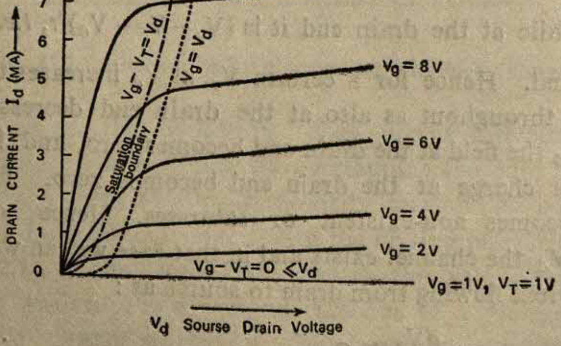


Fig. 11.2.2. *n*-channel enhancement mode transistor output characteristics.

For transconductance we have for equation (11.2.4)

$$g_m = \left.\frac{\delta I_d}{\delta V_g}\right)_{V_d} = \beta V_d \quad \cdots \quad (11.2.7)$$

which increases linearly with V_d till saturation ;
at saturation this is

$$g_{ms} = \beta(V_g - V_T) \qquad \cdots \quad (11.2.8)$$

showing that the saturation region g_m increases linearly with V_g, the gate voltage. The output conductance from (11.2.4) is

$$g_d = \left.\frac{\delta I_d}{\delta V_d}\right)_{V_g} = \beta\,[(V_g - V_T) - V_d] \qquad \cdots \quad (11.2.9)$$

which at saturation is zero in this oversimplified theory. When the channel saturates and, thereafter, increase of V_d causes the pinch-off width of the channel to increase while on either side of this region mobile carrier concentration exists, the field in the pinched-off region increases and the current is mainly controlled by this region. This situation may be compared to an open base swept bipolar transistor or to the saturation region of the velocity field characteristic where near ohms law characteristic is well-known and the output resistance is of finite high value of the order of a few hundred kilo-ohms for high resistivity substrates and long channel devices.

The input impedance of this transistor can be of the order of 10^{14} ohms and entirely depends on how good the insulator leakage reduction has been achieved. The figure of merit or the gain bandwidth product can be defined as the transconductance 'g' divided by the gate capacitance $\left(\frac{\varepsilon\varepsilon_0 WL}{t}\right)$, and is

$$\frac{g_m}{C_g} = \frac{\mu}{L^2}(V_g - V_T) \qquad \cdots \quad (11.2.10)$$

So the frequency performance is proportional directly to the carrier mobility on the surface and inversely to the square of the channel length.

Since the g_m of the device is low, 3 to 5 m mho, it does not find much application as amplifier. It is most imoportant for its very high input impedance and used as mixer, squarer in analog applications and because a MOS transistor used as a resistor occupies about only 2 to 5 per cent space on chip than a conventional resistor as used in article (10.6) so it has found ample application as load in load driver pairs apart from wide spread application in dynamic digital circuits.

11.3 Output and transfer characteristics :

A MOS transistor can work in either enhancement mode where the channel charge strength is enhanced or in depletion mode where the channel charge strength is decreased or in both modes depending on the value of the voltage V_T which is fixed for a certain device. On the other hand, a MOS transistor may have been built on *p*-substrate which will have *n*-channel while if it is built on *n*-substrate it will be a *p*-channel transistor. Considering all these, there can be four types of MOS transistor, the transfer characteristics of some of which are shown in fig. (11.3.1).

In an *n*-channel enhancement mode transistor at $V_g = 0$ no channel exists and the created channel is N-type hence built on *p*-substrate and V_T is positive as shown in fig. (11.2.2). The source is an N-type doping supplying electrons to the channel created by the application of a positive gate voltage as has been discussed. The circuit schematic is a shown in fig. 11.3.1(a). To start with there was no channel indicated by broken line between drain to source.

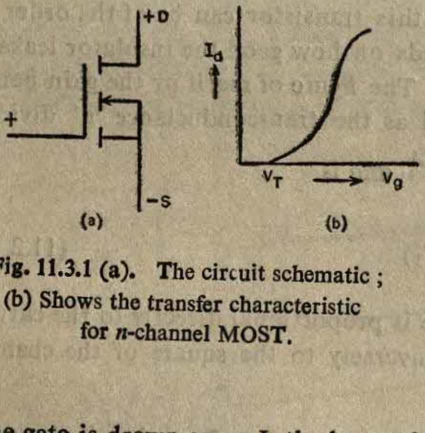


Fig. 11.3.1 (a). The circuit schematic ; (b) Shows the transfer characteristic for *n*-channel MOST.

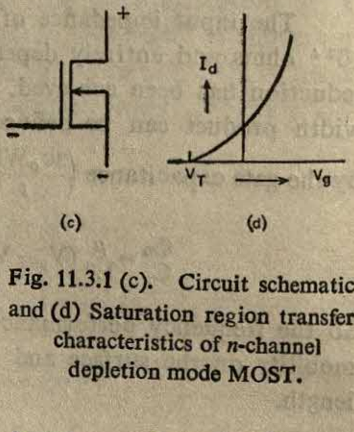


Fig. 11.3.1 (c). Circuit schematic and (d) Saturation region transfer characteristics of *n*-channel depletion mode MOST.

The gate is drawn as an L the knee of which appears near the source. The substrate arrow is pointed in the forward conduction direction of the channel substrate *p-n* junction and hence can be connected to the most negative point ; here the source, thus back-biasing this junction. Drain is collecting electrons ; hence positive with respect to source. V_T is positive and larger gate voltage gives larger current, hence the saturation region transfer characteristic is shown in fig. 11.3.1(b).

For an *n*-channel depletion mode MOSFET at $V_g = 0$, a channel exists, indicated by bold line between source and drain. Gate Knee occurring near the source that supplies electrons to the channel ; hence negative with respect to drain. The gate causes depletion of channel charge ; hence negative with respect to source. A certain negative voltage on gate V_T is needed to pinch-off the channel as shown in fig. 11.3.1(d).

If power dissipation permits this FET can also be used in the enhancement mode by making the gate positive with respect to source as shown in the output characteristics of fig. 11.3.1(e).

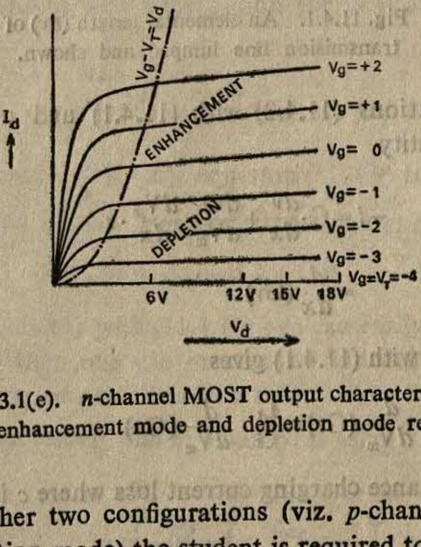


Fig. 11.3.1(e). *n*-channel MOST output characteristics for both enhancement mode and depletion mode regions.

For the other two configurations (viz. *p*-channel enhancement mode and depletion mode) the student is required to plot the characteristics along the lines indicated above.

11.4 Small signal theory :

The channel region of the MOSFET can be regarded as a distributed transmission line of which an elemental section is shown in Fig. (11.4.1) for which

$$I_d\, R dx = -dV_x$$

or

$$I_d = -G \cdot \frac{dVx}{dx} \qquad \cdots \quad (11.4.1)$$

where $G = \frac{1}{R}$ is per unit length conductance of the line at $x = x_1$ say and from equation (11.2.3*a*) is $\varepsilon\varepsilon_o\mu\, W(V_g - V_T - V_x)/t$. Therefore,

$$(Id + i) = -\left[G + \left(\frac{dG}{d_x}\right)_{x=x_1} v\right]\frac{d}{dx}(V_x + v) \quad \cdots \quad (11.4.2)$$

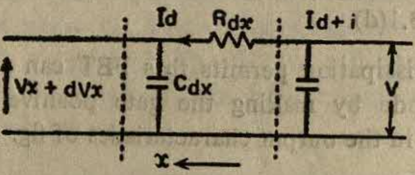


Fig. 11.4.1. An elemental length (δx) of transmision line lumped and shown.

From equations (11.4.2) with (11.4.1) and neglecting second order small quantity

$$-i = G\cdot\frac{dv}{dx} + \frac{dG}{dV_x}\cdot\frac{dV_x}{dx}\cdot v$$

$$= \frac{d}{dx}(Gv) \quad \cdots \quad (11.4.3)$$

equation (11.4.3) with (11.4.1) gives

$$-i = \frac{dV_x}{dx}\cdot\frac{d}{dV_x}(Gv) = \frac{I_d}{G}\cdot\frac{d}{dV_x}(Gv) \quad \cdots \quad (11.4.4)$$

Again for capacitance charging current loss where *c* is per unit length capacitance

$$\frac{di}{dx} = jwcv \quad \cdots \quad (11.4.5)$$

$$= \frac{di}{dV_x}\cdot\frac{dV_x}{dx}$$

Hence using (11.4.1) with (11.4.5)

$$-\frac{di}{dV_x} = j\frac{\omega c}{I_d}(Gv) \quad \cdots \quad (11.4.6)$$

From (11.4.6)

$$-\frac{d^2 i}{dV_x^2} = j\frac{\omega c}{I_d}\frac{d}{dV_x}(Gv)$$

which with equation (11.4.4) gives

$$-\frac{d^2 i}{dV_x^2} = j\frac{\omega c}{I_d^2}\,Gi \qquad \cdots \quad (11.4.7)$$

Substituting for G from article (11.2) and using for $(\varepsilon_s\varepsilon_0 W/t) = c$

$$-\frac{d^2 i}{dV_x^2} = j\frac{\omega c^2\mu}{I_d^2}(V_g - V_T - V_x)\,i \qquad \cdots \quad (11.4.8)$$

changing variable for $(V_g - V_T - V_x) = -u$ and we have from (11.4.8)

$$\frac{d^2 i}{d^2\omega} = j\frac{\omega c^2\mu}{I_d^2}\,ui = jK^2\,ui \qquad \cdots \quad (11.4.9)$$

where $\quad K^2 = (\omega c^2\mu)/I_d^2$.

From equation (11.4.6) with substitution for G as above and change of variable

$$\frac{di}{du} = j\frac{\omega c^2\mu}{I_d}\,uv \qquad \cdots \quad (11.4.10)$$

with transformation $z = \frac{2}{3}K(ju)^{\frac{3}{2}}$ equation (11.4.9) can be reduced to Bessels equation of $\frac{1}{3}$ order. Hence the solution can be written as

$$i = \sqrt{u}[AJ_{\frac{1}{3}}(z) + B\,J_{-\frac{1}{3}}(z)] \qquad \cdots \quad (11.4.11)$$

From equation (11.4.11) with (11.4.10) one can easily calculate v as a function of u and then one can proceed to find out the equivalent circuit along the lines as described in article (9.3).

The equivalent circuit that is obtained is exactly similar in form

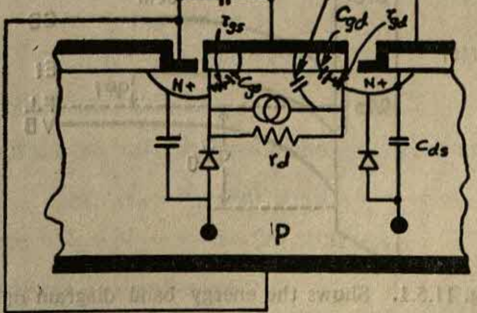


Fig. 11.4.2. Shows the origin of the elements of the equivalent circuit

to that shown in fig. (9.3.3) of article (9.3) and it is instructive to see figs. (11.4.2) and (11.4.3) for the origin of the components.

Note that the gate to channel capacitance C_{ch} gets divided into two parts one adding to C_{gs} and the other to C_{gd}. The source end *p-n* junction diode capacitance is short because substrate is connected to source. The input lead capacitance C_{gss} and the drain diode capacitance C_{ds} remain. For low frequency work, because of the high input impedance, a much simpler equivalent circuit, as shown in fig. (11.4.3) is adequate. For very high frequencies the lead inductances and stray capacitances should be taken into account and the transistor parameters should be measured or collected from the manufacturer for good design work.

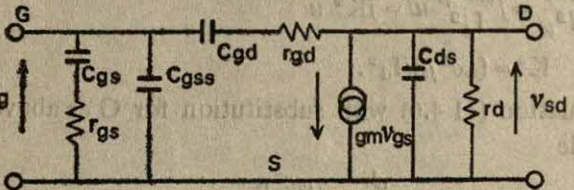


Fig. 11.4.3. Common source equivalent circuit of MOST.

11.5 Semiconductor charge and Potential in MOS :

In order to understand the MOS as capacitor element and its device potential we need to examine the charge distribution and potential variations within the semiconductor under the metal gate

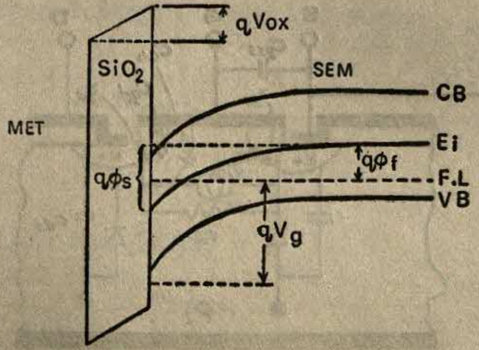


Fig. 11.5.1. Shows the energy band diagram of a *p*-substrate MOS under equilibrium condition after application of a gate voltage V_g with respect to substrate.

region of the MOS. For convenience of description the energy band diagram of a MOS on *p*-substrate is shown in fig. (11.5.1) when a positive gate voltage V_g has been applied to the metal gate with respect to the substrate. In this section we intend to refer to concentrations of carriers in terms of potentials measured with respect to intrinsic Fermi potential ϕ_i where $-q\phi_i = E_i$, when the concentration deep within the bulk can be written as

$$p = n_i \exp(\beta\phi_F) \text{ and } n = n_i \exp(-\beta\phi_F) \quad \cdots \quad (11.5.1)$$

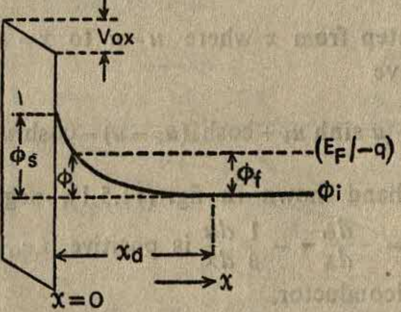


Fig. 11.5.2. Shows the potential diagram corresponding to fig. (11.5.1).

where $\beta = (q/kT)$, *i.e.*, for an N-type semiconductor ϕ_F is negative. Similarly, for a location in the space charge region where the departure of the intrinsic potential level is ϕ from its value deep within the bulk the concentrations for the situation shown in fig. (11.5.1) can be written as

$$\left.\begin{aligned} p &= n_i \exp \beta(\phi_F - \phi) \\ n &= n_i \exp - \beta(\phi_F - \phi) \end{aligned}\right\} \quad \cdots \quad (11.5.2)$$

under the equilibrium situation.

Deep within the bulk the space charge is zero and so

$$p\text{-}n = N_a - N_d \text{ and } \phi = 0$$

$$\text{hence} \quad N_a - N_d = p - n = 2n_i \sinh \beta\phi_F$$

While in the space charge region

$$\begin{aligned} \rho &= q(p - n - N_a + N_d) \\ &= 2n_i q\,[\sinh \beta(\phi_F - \phi) - \sinh \beta\,\phi_F] \quad \cdots \quad (11.5.3) \end{aligned}$$

Hence for the space charge region the Poisson's equation may be written with $\beta\phi = u$ notation as

$$\frac{d^2\beta\phi}{dx^2} = -\frac{2qn_i\beta}{\varepsilon_s\varepsilon_0}\left[\sinh(u_F - u) - \sinh u_F\right] \quad \cdots \quad (11.5.4)$$

defining $\sqrt{\frac{\varepsilon_s\varepsilon_0}{2qn_i\beta}} = L$, where L is the Debye length.

We have from (11.5.4)

$$L^2 \frac{du^2}{dx^2} = \left[\sinh u_F - \sinh(u_F - u)\right] \quad \cdots \quad (11.5.5)$$

Integrating one step from x where $u = u$ to $x = \infty$ where $u = 0$ and $(du/dx) = 0$ we have

$$\frac{L}{\sqrt{2}}\frac{du}{dx} = \pm\left[u \sinh u_F + \cosh(u_F - u) - \cosh u_F\right]^{\frac{1}{2}} \quad \cdots \quad (11.5.6)$$

For the case in hand shown in fig. (11.5.1), negative sign applies since the field $E = -\frac{d\phi}{dx} = -\frac{1}{\beta}\frac{du}{dx}$ is positive, *i.e.*, directed from the oxide to the semiconductor.

From equation (11.5.6) the semiconductor surface field is obtained with $\phi = \phi_s$ and $u = u_s$ as

$$E_s = \frac{\sqrt{2}}{\beta L}\left[u_s \sinh u_F + \cosh(u_F - u_s) - \cosh u_F\right]^{\frac{1}{2}} \quad \cdots \quad (11.5.7)$$

Hence the total charge per unit area within the semiconductor by using Gauss' law, since deep inside the semiconductor the field is zero, is given by

$$Q_s = -\varepsilon_0\varepsilon_s E_s = -\frac{\sqrt{2}\varepsilon_s\varepsilon_0}{\beta L}\left[u_s \sinh u_F + \cosh(u_F - u_s) - \cosh u_F\right]^{\frac{1}{2}} \quad \cdots \quad (11.5.8)$$

For the *p*-type substrate under discussion, the potential ϕ variation is shown in fig. (11.5.2). The space charge (Q_s/q) versus ϕ_s can be plotted for different extent of doping, *i.e.*, for different values of ϕ_F from equation (11.5.8) and is shown in fig. (11.5.3). It is to be noted from this figure that strong change of minority carrier concentration occurs at about $\phi_s = 2\phi_F$; hence we take the surface potential on start of strong inversion as $2\phi_F$ for computational purposes. This can be given a plausibility argument that in the bulk there was a certain concentration of majority carriers given by equation (11.5.1), due to band bending, whatever may be the reason, the intrinsic Fermi potential need to depart by ϕ_F to turn the point in question to intrinsic and a further amount of ϕ_F to convert this point equally strong of opposite conductivity type.

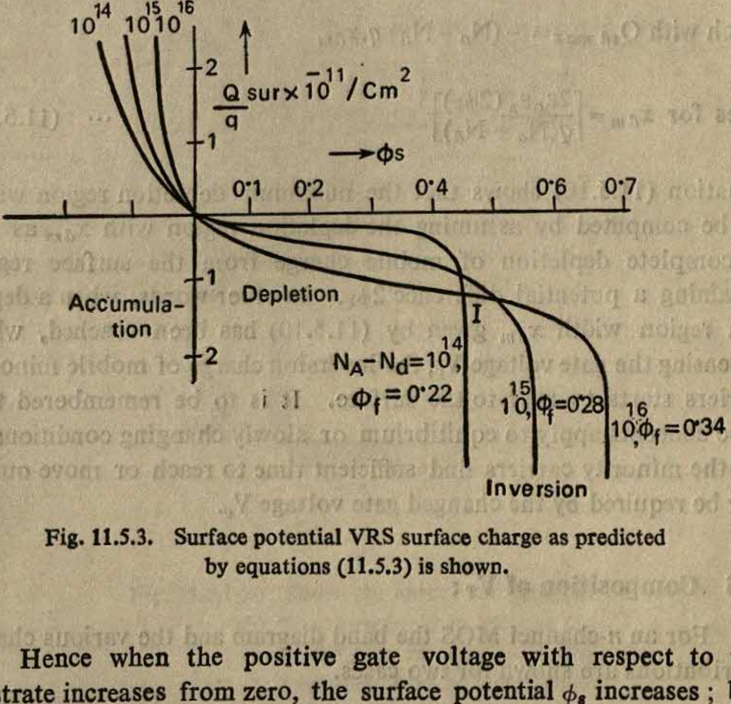


Fig. 11.5.3. Surface potential VRS surface charge as predicted by equations (11.5.3) is shown.

Hence when the positive gate voltage with respect to the substrate increases from zero, the surface potential ϕ_s increases ; but on strong inversion that occurs at $\phi_s = 2\phi_F$, the surface potential cannot change any further since mobile minority charges of large magnitude rushes to the surface to quench any excess field that may be attempted to be set up within the semiconductor from outside application of voltage. This can reasonably be viewed as a failure on the part of the surface recombination rate, that cannot cope up with the on-rush of minority carriers at this level of field, to annihilate them by recombination.

Hence with increase of gate voltage if ϕ_s does not change so the depletion region width will not also change. That is x_d the depletion region width in a MOS, will have a maximum value under *equilibrium* conditions as x_{dm}. It is to be noted that it is not a maximum in the algebraic sense rather an approximation. From equation (11.5.8) with the onset of inversion $\phi_s = 2\phi_F$ and using $(N_a - N_d) = 2n_i \sinh u_f$ we have for the per unit area fixed surface charge at the start of inversion as

$$Q_{sd\ max} = -[2\varepsilon_0\varepsilon_s q (N_a - N_d)(2\phi_F)]^{\frac{1}{2}} \quad \cdots \quad (11.5.9)$$

which with $Q_{sd\ max} = -(N_a - N_d)\, q.x_{dm}$

$$\text{gives for } x_{dm} = \left[\frac{2\varepsilon_0\varepsilon_s (2\phi_F)}{q(N_a - N_d)}\right]^{\frac{1}{2}} \quad \cdots \quad (11.5.10)$$

Equation (11.5.10) shows that the maximum depletion region width can be computed by assuming the depletion region with x_{dm} as due to complete depletion of mobile charge from the surface region sustaining a potential difference $2\phi_F$. In other words, when a depletion region width x_{dm} given by (11.5.10) has been reached, while increasing the gate voltage V_g, the inversion charge of mobile minority carriers starts to rush to the surface. It is to be remembered that these concepts apply to equilibrium or slowly changing conditions so that the minority carriers find sufficient time to reach or move out as may be required by the changed gate voltage V_g.

11.6 Composition of V_T :

For an *n*-channel MOS the band diagram and the various charge distributions are shown for two cases.

Case 1 : Where the gate voltage is zero.

Case 2 : When $x_d = x_{dm}$ and a gate voltage larger than required to produce inversion has been applied.

The various charges and energy differences that appear in the figs. (11.6.1) and (11.6.2) are :

$q\,\phi_{Mox}$ the metal to oxide barrier energy, depends on the metal and the particular oxide in question. As for example, photoelectric and C-V measurements suggest for $Al - SiO_2$ system $q\phi_{Mox} = 3.2$ ev, for $Au - SiO_2$ $q\,\phi_{Mox} = 4.1$ ev and for $Ni - SiO_2$ system $q\,\phi_{Mox} = 3.7$ ev and so on.

$q\phi_{Sox}$ is the energy associated with Si-SiO$_2$ barrier. Measurements made with various metals with Si-SiO$_2$ interface indicate that

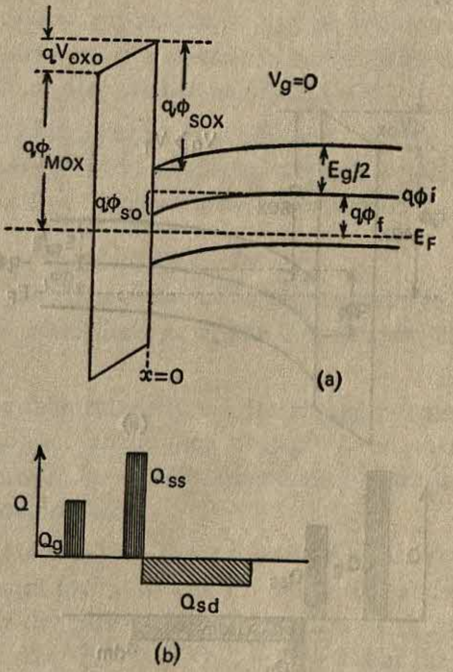


Fig. 11.6.1 (a) Shows the energy band diagram; (b) The corresponding charges Q_g at the gate metal, Q_{ss} at the oxide-sem interface and Q_{sd} the semiconductor depletion region charge at zero bias between metsem.

Silicon-Silicon dioxide barrier energy is 3.25 ev, and is approximately independent of conductivity of Si or its conductivity type and its orientation.

qV_{oxo} is the potential energy difference across the insulator under zero voltage gate to substrate bias condition and depends on the technological process step details and success.

$q\phi_{so}$ is the zero bias surface potential energy with reference to the bulk, hence also depends on the technological process step details and their success.

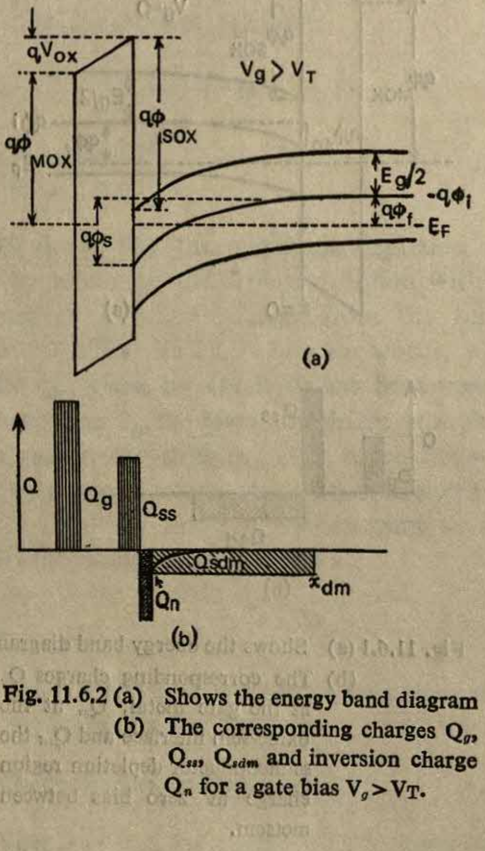


Fig. 11.6.2 (a) Shows the energy band diagram
(b) The corresponding charges Q_g, Q_{ss}, Q_{sdm} and inversion charge Q_n for a gate bias $V_g > V_T$.

$q\phi_F$ is the energy difference between the intrinsic Fermi level and Fermi level in the bulk far away from the contact region hence independent of bias conditions (c.f. equation 11.5.1).

$q\phi_s$ is the surface potential energy under a gate bias V_g with respect to the bulk crystal.

qV_{ox} is the potential energy difference across the insulator for a gate bias voltage V_g.

Q_{ss} in figs. (11.6.1) and (11.6.2) is the fixed normally positive surface charge density per unit area present in the silicon dioxide very close to the semiconductor silicon interface. This charge depends on the surface states as also on the technological process details and depends on the crystal orientation. Any mobile ionic charge in the insulator SiO_2 is assumed not to exist in contributing to Q_{ss}. Hence Q_{ss} is independent of bias conditions of the MOS.

Q_g in fig. (11.6.1a) is the per unit area charge needed on the metal gate to account for total charge neutrality under equilibrium conditions at zero bias. Hence under the condition

$$Q_g + Q_{ss} + Q_{sd} = 0.$$

Q_{sdm} is the maximum depletion region charge per unit area under equilibrium conditions at a gate to substrate bias voltage V_g such that $\phi_s = 2\phi_f$.

Q_n is the mobile minority carrier charge per unit area under inversion condition at an applied voltage V_g on metal gate with respect to substrate under equilibrium condition and will henceforth be called the inversion charge.

Q_g in fig. (11.6.2a) is the per unit area charge needed on the metal gate to account for total charge neutrality under equilibrium conditions at a positive gate voltage V_g applied with respect to the substrate. Here also $Q_g + Q_{ss} + Q_{sdm} + Q_n = 0$ but the values of Q_g in the two figures are different.

From Gauss' law in either of the figs. (11.6.1a) and (11.6.2a) the potential difference across the insulator is (Q_g/C_{ox}) where $C_{ox} = \left(\frac{\varepsilon_{in}\varepsilon_o}{t_{ox}}\right)$ and t_{ox} is the insulator thickness.

From fig. (11.6.1a) we have,

$$\phi_{Mox} + V_{oxo} = \phi_{sox} + \frac{E_g}{2q} - \phi_{so} + \phi_f \qquad \cdots \qquad (11.6.1)$$

Defining for the particular metal-insulator-semiconductor combination, the constant

$$\phi_{MS} = \phi_{Mox} - \phi_{sox} - \frac{E_g}{2q} - \phi_f \qquad \cdots \qquad (11.6.2)$$

we have from (11.6.1)

$$\phi_{MS} = -(\phi_{so} + V_{oxo}) \qquad \cdots \qquad (11.6.3)$$

Considering the equilibrium situation depicted in fig. (11.6.2a), that the variable quantities in equation (11.6.1) are ϕ_{so} and V_{oxo} only and in both of the figures these two voltages being additive and in the same sense of deviation, we have for V_g

$$V_g=(\phi_x+V_{ox})-(\phi_{so}+V_{oxo}) \quad \cdots \quad (11.6.4)$$

which with (11.6.3) gives

$$V_g=\phi_{MS}+V_{ox}+\phi_s \quad \cdots \quad (11.6.5)$$

Compare this with equation (11.6.3) to find its general character regarding variation of ϕ_s and V_{ox} while Gauss' law gives always

$$V_{ox}=Q_g/C_{ox} \quad \cdots \quad (11.6.6)$$

and under inversion conditions,

$$Q_g+Q_{ss}+Q_{sdm}+Q_n=0 \quad \cdots \quad (11.6.7)$$

while prior to inversion $Q_{sd} < Q_{sdm}$

and $$Q_g+Q_{ss}+Q_{sd}=0 \quad \cdots \quad (11.6.8)$$

Equations (11.6.5) to (11.6.8) describe the charges and potential variables in a MOS structure.

From equation (11.6.5) at the onset of inversion

$$V_g=V_T=\phi_{MS}+Q_g/C_{ox}+2\phi_f$$

and since $Q_n=0$, $Q_{sd}=Q_{sdm}$

$$V_T=\phi_{MS}-\frac{(Q_{ss}+Q_{sdm})}{C_{ox}}+2\phi_f \quad \cdots \quad (11.6.9)$$

while $Q_{sdm}=\pm\sqrt{2\varepsilon_0\varepsilon_s q(N_a-N_d)\,2\phi_f} \quad \cdots \quad (11.6.10)$

minus sign being applicable to *p*-type substrate while plus sign for *n*-substrate where though ϕ_f is negative (N_a-N_d) is also negative.

Hence from equations (11.6.9) and (11.6.10) the threshold voltage of a *p*-substrate *n*-MOS is given by

$$V_T=\left(\phi_{MS}+\frac{1}{C_{ox}}\sqrt{2\varepsilon_0\varepsilon_s q(N_a-N_d)(2\phi_f)}-\frac{Q_{ss}}{C_{ox}}+2\phi_f\right) \quad \cdots \quad (11.6.11)$$

It is to be remembered for *n*-substrate (N_a-N_d) is negative so also ϕ_f.

Because of the presence of Q_{ss} even under zero applied gate voltage conditions there can have bent bands and to take account of this band bending equation (11.6.5) can be written with equation (11.6.8) as

$$V_g = \phi_{MS} - \frac{Q_{ss}}{C_{ox}} - \frac{Q_{sd}}{C_{ox}} + \phi_s = V_{fb} - \frac{Q_{sd}}{C_{ox}} + \phi_s \qquad \cdots \quad (11.6.12)$$

where the flat band voltage

$$V_{fb} = \phi_{MS} - \frac{Q_{ss}}{C_{ox}} \qquad \cdots \quad (11.6.13)$$

is the gate voltage needed to make $\phi_s = 0 = Q_{sd}$ i.e., to make the bands flat.

Using equations (11.6.13) equation (11.6.11) gives for the threshold voltage of a MOS

$$V_T = V_{fb} + 2\phi_f \pm \frac{1}{C_{ox}} \sqrt{4\varepsilon_0 \varepsilon_s q(N_a - N_d)\phi_f} \qquad \cdots \quad (11.6.14)$$

The negative sign applies for *n*-type substrate.

It is to be noted that the threshold voltage depends on the flat band voltage *i.e.*, the surface conditions described by Q_{ss}, the metal oxide system described by ϕ_{MS}(cf. equation 11.6.2), the doping density described by $(N_a - N_d)$ also ϕ_f and the oxide capacitance as determined by the oxide thickness and its permittivity.

11.7 MOS Capacitance and its equivalent circuit

The substrate to gate capacitance of the MOS structure depends on the d.c. applied voltage on the gate as also on the frequency. The d.c. applied voltage can cause either accumulation, depletion or even inversion of the oxide semiconductor surface layer, while in inversion condition only, the minority carriers are called into play in large numbers. In accumulation majority carriers are involved to respond to the a.c. voltage fluctuation required for the measurement of the capacitance. In the depletion region also the removal and calling back of the majority carriers are involved mainly. While in inversion on top of removal and call of the majority carriers for the depletion region, the rush of minority carriers to the surface from the depletion region obtained from thermal generation or from the neutral region on all sides of the depletion region occurs, which, the surface being unable to completely anihilate by recombination; shares a portion of the a.c. field. Hence, the capacitance versus applied gate voltage is frequency dependent as shown in fig. (11.7.1).

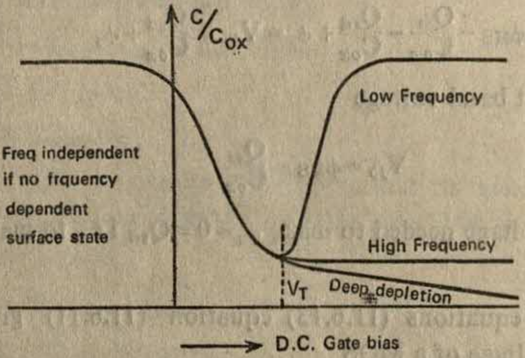


Fig. 11.7.1. MOS capacitance versus substrate gate voltage of *p*-substrate.

For an examination of the detailed nature of variation of this capacitance, we write

$$C^{-1} = \frac{dV_g}{dQ_g} \qquad \cdots \quad (11.7.1)$$

which with equation (11.6.5) can be written as

$$C^{-1} = \frac{dV_{ox}}{dQ_g} + \frac{d\phi_s}{dQ_g} \qquad \cdots \quad (11.7.2)$$

Equation (11.6.7) can be re-written as

$$|Q_g| = |Q_{ss} + Q_{sdm} + Q_n| \qquad \cdots \quad (11.7.3)$$

Now define $\frac{d|Q_g|}{dV_{ox}} = C_{ox}$ the oxide capacitance ... (11.7.4a)

$$\frac{d|Q_{ss}|}{d\phi_s} = C_{ss} \qquad \cdots \quad (11.7.4b)$$

if Q_{ss} or its part is variable

$\frac{\varepsilon_0 \varepsilon_s}{x_{dmax}} = \frac{d|Q_{sdm}|}{d\phi_s} = C_{sdm}$ if Q_{sdm} varies due to ϕ_s variation

... (11.7.4c)

and $\frac{d|Q_n|}{d\phi_s} = C_n$ minority carrier capacitance. ... (11.7.4d)

Equation (11.7.2) can be written as

$$C^{-1}=C_{ox}^{-1}+\left[\frac{d|Q_{\theta}|}{d\phi_s}\right]^{-1}$$

which with equations (11.7.3) and (11.7.4) becomes

$$C^{-1}=C_{ox}^{-1}+(C_{ss}+C_{sdm}+C_n)^{-1} \quad \cdots \quad (11.7.5)$$

Under accumulation conditions only the majority carrier fluctuation due to the a.c. applied voltage is of major importance since the original minority carriers are too few to be counted. While the majority carriers can respond to the fluctuating voltage with the relaxation time $\tau=\varepsilon_s\varepsilon_0/\sigma$ which is of the order of picoseconds or lower depending on the doping level of the substrate. Hence majority carrier charge changes will be independent of frequency. Moreover, the terms C_{sdm} and C_n in the second term of the right-hand side of equation (11.7.5) does not concern us here since there is neither depletion charge nor inversion (minority) charge involved in the accumulation region. But on the surface, there can be surface states which can capture or release majority carriers as the Fermi level moves with a.c. applied voltage. In that case against C_{ss} we can use a suitable capacitance C_{st} defined by $C_{st}=\frac{d|Q_{st}|}{d\phi_s}$ where $Q_{ss}=Q_{sso}+Q_{st}$, and $C_{ss}=\frac{dQ_{ss}}{d\phi_s}=C_{st}$; Q_{sso} being the constant part, hence under accumulation region the gate to substrate capacitance is given by

$$C^{-1}=C_{ox}^{-1}+(C_{st})^{-1} \quad \cdots \quad (11.7.6)$$

which we shall see later to depend on surface state's speed of particle exchange. In ideal MOS capacitors voltage variable surface state charge density is assumed to be low enough so that Q_{ss} is practically constant in that case the capacitance in accumulation region is given by

$$C=C_{ox} \quad \cdots \quad (11.7.7)$$

and is independent of frequency as shown in fig. (11.7.1).

Under depletion conditions, there is no inversion charge Q_n ; hence $C_n=0$ for ideal MOS, Q_{ss} is fixed hence $C_{ss}=0$, and we consider in this case $C_{sd}=\frac{d|Q_{sd}|}{d\phi_s}$ against C_{sdm} of equation (11.7.5) since the onset of inversion has not yet been reached, when we have from equation (11.7.5) for ideal MOS capacitance.

$$C^{-1} = C_{ox}^{-1} + C_{sdm}^{-1} \quad \cdots \quad (11.7.8)$$

Here again the movement of majority carriers are responsible for the charge depletion and their coverage hence the capacitance in this region given by equation (11.7.8) is independent of frequency as shown in fig. (11.7.1). If, of course, the capacitance is non-ideal then we should take account of surface state charge variation as was done in the accumulation region and the capacitance in that case will be given from (11.7.5) with $C_{ss} = \dfrac{dQ_{ss}}{d\phi_s} = \dfrac{dQ_{st}}{d\phi_s} = C_{st}$

$$\text{as } C^{-1} = C_{ox}^{-1} + (C_{st} + C_{sd})^{-1} \quad \cdots \quad (11.7.9)$$

Within the inversion regim again equation (11.7.5) is used assuming ideal case situation of $C_{ss} = 0$ i.e., the surface state charge is fixed. For very low frequencies, the minority carriers find enough time to reach the surface in response to the field increment which, in turn, quenches this incremental field, thus inhibiting excess depletion necessary to accommodate such a.c. fields. Hence we have

$$C_{sdm} = \frac{d|Q_{sdm}|}{d\phi_s} = 0 \text{ while } C_n = \frac{d|Q_n|}{d\phi_s}$$

tends to very large value. Hence the post inversion low frequency MOS capacitance can be written from equation (11.7.5) as

$$C^{-1} = C_{ox}^{-1} + C_n^{-1} \simeq C_{ox}^{-1} \quad \cdots \quad (11.7.10)$$

which varies with d.c. gate bias shown in fig. (11.7.1).

At very high frequencies, on the other hand, the minority carriers fail to respond to the incremental field because their generation recombination life time is too large and hence they take longer time to respond, so the depletion region majority charge moves out to accommodate the excess a.c. field. Hence, we have $\dfrac{d|Q_{sdm}|}{d\phi_s} = C_{sdm} = (\varepsilon_s\varepsilon_0/x_{dmax})$ and $C_n = \dfrac{d|Q_n|}{d\phi_s}$ tending to zero. Using these with Q_{ss} fixed equation (11.7.5) gives for very high frequencies post inversion capacitance as

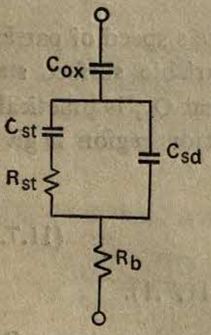


Fig. 11.7.2. Shows the equivalent circuit model of MOS capacitor.

$$C^{-1} = C_{ox}^{-1} + C_{sdm}^{-1} \quad \cdots \quad (11.7.8)$$

Considering the above discussions and that there would have some amount of bulk resistance the MOS capacitance in depletion be represented by fig. (11.7.2) where for high frequency, C_{sd} is to be replaced by C_{sdm} after inversion and C_{st} and R_{st} represent the frequency dependent surface states.

11.8 MOS Capacitance under deep depletion

In the previous article we have seen that the ideal high frequency capacitance in post-inversion condition attains a constant value given by equation (11.7.8). This is because the minority carriers fail to take effective part in quenching the high frequency field changes because of their comparatively larger generation times. Alternatively, if the generated minority carriers are continuously withdrawn as it happens in a reverse biased *p-n* junction or may happen in a MOS capacitor having a poor (leaky) oxide or in the case of application of a step voltage bringing the MOS from accumulation to inversion, in all these cases minority carriers will not be adequately available for responding to the incremental voltage changes ; hence the depletion layer width will increase beyond its x_{dmax} value and the capacitance is expected to decrease further than the high frequency value given by equation (11.7.8). To obtain a quantitative estimate of this pulse voltage dependence, we again assume an ideal MOS i.e., Q_{ss} is fixed hence $C_{st} = 0$ while the depletion region is completely depleted of mobile carriers.

Hence

$$\phi_s = \int_0^{x_d} \frac{qN_a x\,dx}{\varepsilon_o \varepsilon_s} = \frac{qN_a x_d^2}{2\varepsilon_s \varepsilon_o} = \frac{qN_a \varepsilon_s \varepsilon_o}{2C_{sd}^2} \quad \cdots \quad (11.8.1)$$

and $$C^{-1} = C_{ox}^{-1} + C_{sd}^{-1} \quad \cdots \quad (11.8.2)$$

whereas from equation (11.6.12)

$$\phi_s = V_g - V_{fb} - \frac{qN_a \varepsilon_s \varepsilon_o}{C_{ox}\, C_{sd}} \quad \cdots \quad (11.8.3)$$

since $Q_{sd} = -qN_a x_d$ and $C_{sd} = (\varepsilon_s \varepsilon_o / x_d)$

From equations (11.8.1) and (11.8.3) eliminating ϕ_s and re-arranging we have

$$(C_{ox}/C_{sd})^2 + 2(C_{ox}/C_{sd}) - \frac{2C_{ox}^2(V_g - V_{fb})}{qN_a \varepsilon_s \varepsilon_o} = 0 \quad \cdots \quad (11.8.4)$$

Writing equation (11.8.2) in the form

$$(C_{ox}/C_{sd})=\left(\frac{C_{ox}}{C}\right)-1=\alpha \qquad \cdots \quad (11.8.5)$$

we have from equations (11.8.4) and (11.8.5)

$$\alpha=-1+[1+(2C_{ox}{}^2\,\overline{V_g-V_{fb}}/qN_a\varepsilon_0\varepsilon_s]^{\frac{1}{2}}=(C_{ox}/C)-1$$

Hence $$\frac{C}{C_{ox}}=[1+(2C_{ox}{}^2\,\overline{V_g-V_{fb}}/\varepsilon_s\varepsilon_0 qN_a)]^{-\frac{1}{2}} \qquad \cdots \quad (11.8.6)$$

This capacitance decrease is depicted in fig. (11.8.1). V_{fb} is normally negative since Q_{sso} is always positive hence $V_{fb} \neq 0$ curve is on the left of $V_{fb}=0$ curve for *p*-substrate MOS.

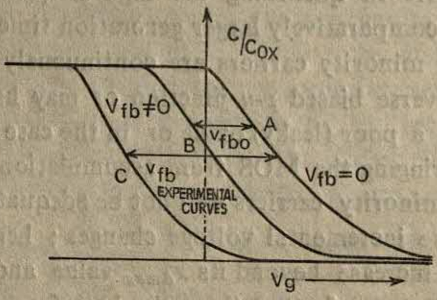


Fig. 11.8.1. Shows theoretical C/C_{ox} vrs. substrate to gate-voltage when surface inversion has failed to set in.

11.9 Surface States and MOS-CV Characteristics

Surface energy seats at the Si-SiO$_2$ interface lying within the band gap can steadily capture from or release charges to the adjacent substrate. These energy seats are called fast surface states. These states are known (c.f. ref. 1) to have pronounced maximum density distribution near the band edges with the minimum at mid-gap in silicon. With change of gate voltage the position of the F.L. with respect to the intrinsic F.L.; hence the mid-gap, at the interface changes thus causing a variation of occupational probability of these states. This causes trapping and subsequent release of appreciable number of free carriers with change of gate-voltage. The origin of these states is to some extent due to disruption of the periodicity of the silicon lattice and excess Si-ion movement into the oxide leaving vacant sites on the Si-surface during oxidation, is seen to increase these fast surface states. The density of these states can be reduced to very low levels by low temperature annealing in inert atmosphere for Al-SiO_2-Si system. It has also been found that the fast surface state density is greatest on (111) plane and smallest on (100) plane of silicon. The states that are near the band centre are responsible for surface recombination and act as scattering centres on the interface thus reducing the mobility of carriers on the silicon surface. In this text the charges in these states have been denoted by Q_{st}.

Another type of always positive surface charge exists in the insulator within 200 Å of the Si-SiO_2 interface whose per unit area value is fixed. The energy states responsible for these charges lie close to the conduction band edge and above. They are known to be donor-like states and are permanently ionised. Hence their occupational probability is not affected by applied voltage or moderate band bending.

It is found that like the fast states the fixed charge density depends on crystal orientation, largest on (111) plane and smallest on (100) plane. Their magnitude depends upon temperature above 250°C and on the ambient atmosphere. Though the fast states and fixed states are identical in their behaviour in some respects their origin are entirely different and they can be independently controlled (c.f. ref. 2). It is known that the last high temperature treatment sets the final Q_{sso} level apart from the others while low Q_{sso} levels are obtained by growing the gate insulator in dry oxygen at a suitable temperature and then subjecting the wafer to an inert high temperature anneal in dry nitrogen.

If the MOS structure is an ideal one, then $Q_{ss} = Q_{sso}$ and there are fixed charges only. Additionally, if V_{fb} the flat band voltage is zero, then the high frequency capacitor during depletion will be given by equation (11.8.6) as shown in fig. (11.8.1). If V_{fb} is not zero, then the curve will shift parallel to itself by the amount V_{fb} to the left as shown by curve B in this figure, since, in general, Q_{sso} is positive.

Assuming that there are no mobile ionic charges in the insulator, that shifts with applied gate-voltage, the experimental curve will exhibit the contribution due to fast surface states also and would be somewhat like curve C. If V_{fb} can be ascertained from the experimental C-V curve and the ideal curve A, then Q_{sso} can be estimated by using the relation $V_{fb} = \phi_{MS} - \frac{Q_{sso}}{C_{ox}}$. For the determination of detailed nature of surface charges the conductance method and temperature variation methods discussed in ref. (1) should be used.

11.10 Effect of back gate bias on V_T

If the source diffusion to substrate junction is reverse biased, for example, the *p*-type substrate is made negative with respect to source then, even if the gate-voltage is not applied, there will have a depletion region on the surface and the generated minority carriers in this depletion region will be drained to the source as it occurs in a reverse biased *p-n* junction diode. If one tries to invert the surface by applying gate bias under this back biased situation, obviously, an extra depletion width has to be created the generation in which will compete over the surface annihilation and drainage process. In other words, the depletion region width under this situation is more than without substrate source reverse bias. Without this back gate bias, so called because the substrate is acting as second gate, a surface potential of $2\phi_f$ is necessary for inversion of surface which gives a maximum per unit area depletion charge Q_{sdm}. Assuming one-sided abrupt junction and a junction reverse bias voltage V_s, the surface potential at the drain end of the gate with respect to the source has to be raised by the same amount by applying positive gate-voltage to overcome the *p-n* junction voltage drop from source to drain end of the gate. This will cause an extra depletion region over that necessary for inversion without back gate bias. Under this situation, if the maximum depletion region charge per unit area is Q_{sdm}', then the depletion region charge with junction reverse bias is related to the charge without back gate bias as

$$Q'_{sdm} = Q_{sdm} \left[\left(|2\phi_f| + |V_s| \right) / 2 |\phi_f| \right]^{\frac{1}{2}} \quad \cdots \quad (11.10.1)$$

Hence the threshold voltage for back gate bias reduces to

$$V_T = V_{fb} + 2\phi_f \pm \left[2\varepsilon_0 \varepsilon_s q N (|2\phi_f| + |V_s|) \right]^{\frac{1}{2}} / C_{ox} \quad \ldots \quad (11.10.2)$$

Plus and minus signs are applicable for *p* and *n* substrates respectively (cf. equations (11.6.9), (11.6.10) and (11.6.14)) where

$N = |N_a - N_d|$. Note that magnitude of V_T increases for enhancement mode transistor while it decreases for depletion mode transistor.

11.11 Effect of Temperature on Current

The current in a MOS transistor can vary due to two causes. Firstly, through variation of mobility (cf. equation 11.2.6) and secondly due to variation of V_T (cf. equation 11.2.5 and 11.6.14). To speak of the variation of V_T we have from equation (11.5.1) $N_d = n_i \exp(-\beta\phi_F)$ while $n_i^2 = N_C N_V \exp(-E_g/kT)$ also the variation of band gap with temperature is known to be of the form $E_g = E_{go} - \alpha T^2/(T+\beta)$ where for Si the values of $\alpha = 7{\cdot}02 \times 10^{-4}$, $\beta = {\cdot}1108$ and $E_{go} = 1{\cdot}16$ (cf. ref.1). The contribution of band gap variation with temperature turns out to be small. Assuming ϕ_{MS} as temperature independent, we can write from (11.6.14) after differentiation and subsequent substitution

$$(\delta V_T/\delta T) = (\delta\phi_f/\delta T)\left(1 + \frac{V_T - V_{fb}}{2\phi_f}\right) \qquad \ldots \quad (11.11.1)$$

while since the variation of exponential term is appreciably large compared to $N_C N_V$ term in n_i,

$$(\delta\phi_f/\delta T) = \frac{1}{T}\left[\frac{E_g}{2q} + \phi_f\right] \qquad \ldots \quad (11.11.2)$$

Hence from (11.11.1) with (11.11.2)

$$(\delta V_T/\delta T) = \frac{1}{T}\left(\frac{E_g}{2_a} + \phi_f\right)\left(1 + \frac{V_T - V_{fb}}{2\phi_f}\right) \qquad \cdots \quad (11.11.3)$$

For an n-substrate doping of 10^{15}/cc, a $V_T = -4{\cdot}1$ and $V_{fb} = -3$ volts the value of $\phi_f = -0{\cdot}29$ and $\left(\frac{V_T}{\delta T}\right) = +0{\cdot}0029$ V/°C. Experiment shows, V_T varies linearly with temperature over useful temperature range having an average temperature coefficient of about this value.

When we think of the variation of mobility since both the impurity scattering and lattice scattering decide the variation of mobility, we can take $\mu \alpha T^n$ where n is decided by predominance of one scattering mechanism over the other. The variation of current I_d with temperature can be written from equation (11.2.5) after rearrangement as

$$(\delta I_d/\delta_T) = I_d\left(\frac{n}{T} - \frac{2}{|V_g - V_T|} \cdot \frac{\delta V_T}{\delta T}\right) \qquad \cdots \quad (11.11.4)$$

Since in silicon *n* is about minus one, for the above-mentioned *n*-substrate MOS with a gate-voltage of $-9{\cdot}1$ volts, the temperature coefficient of I_d is negative and 0·22% around room temperature. Note that the effect of variation of mobility is far in excess over that due to the variation of V_T with temperature in this calculation.

11.12 Standard (111) Process of MOST

The method utilised in fabricating *p*-channel enhancement mode MOST on the (111) crystal plane as used for bipolar transistor fabrication is known as the standard (111) MOS process. In this, one starts with an *n*-substrate in (111) orientation.

Step (i) : The slice is lapped, polished, cleaned and washed.

Step (ii) : A layer of oxide of the order of 7500 Å is grown over it as shown in fig. 11.12.1(a).

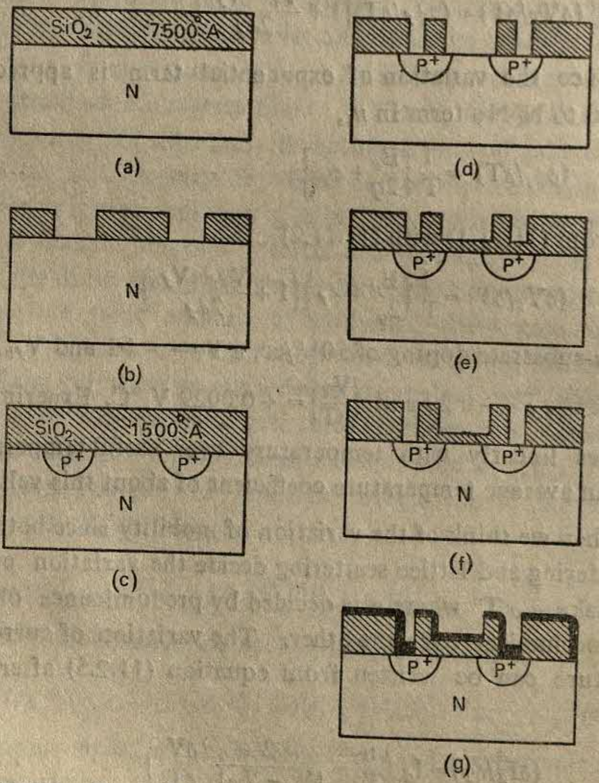


Fig. 11.12.1. Shows the processing steps of standard (111) process

Step (iii) : By using a standard photolithographic step with a mask the source and drain regions are defined and oxide from these locations are etched out as shown in fig. 11.12.1(b).

Step (iv) : The wafer is then placed in diffusion furnace at 1100°C for deposition and diffusion of boron in the exposed source drain regions. Continuing this step or using a separate oxidation step, one obtains a thick oxide of the order of 15000 Å over the whole of the slice including the P-diffused regions as shown in fig. 11.12.1(c).

Step (v) : By using a standard photolithographic step with another suitable mask the source drain and gate openings are defined and subsequently etched out when one obtains as in fig. 11.12.1(d). Note smaller sized openings in this figure.

Step (vi) : After cleaning, the water is placed in oxidation furnace and a layer of oxide of the order of 1200 Å is grown in dry oxygen atmosphere since this step gives the gate oxide whose quality is required to be good. After this oxidation the wafer is shown in fig. 11.12.1(e).

Step (vii) : This is the third masking step which is a standard photolithographic step with a suitable mask like the mask used in step (iii) above but with reduced openings corresponding to the mask used in step (v). In this step after defining the openings the source and drain region oxides are etched out when one obtains the slice as shown in fig. 11.12.1(f).

Step (viii) : In this step Al metallisation is made over the whole slice when one obtains the metallised discrete device as shown in fig. 11.12.1(g). The metallisation evaporation is allowed to impinge at an inclined angle to give breaks as shown in this figure. The gate. source and drain metallisations connect to their bonding pads respectively.

Note that in the standard (111) process described above, only three masking, barely two oxidation, one diffusion and one metallisation steps were necessary. If instead, MOS circuits were fabricated, then one would have started with a *p*-substrate and the third step above would have been a masking step for defining isolation islands where *n*-diffusion would have been made to realise *n*-islands, and an oxidation would have followed corresponding to the present step (ii) discussed above. Again, in the end would have followed another masking step for inter-connection pattern etching. Thus, in MOS circuit fabrication one needs five masking steps, at least three oxidation and two diffusion steps. Thus, we see that at least one diffusion and one oxidation step is reduced in MOS circuit fabrication than bipolar circuit fabrication cf. Art. 10.6 Step (6). This decrease of the number of process steps obviously increases the percentage yield and gives one count in favour of low cost of MOS technology. Secondly, the MOST can be used as a resistor by connecting the gate to drain. Such a resistor needs much less silicon area than a conventional bipolar resistor. Since a typical 100 Ohm/sq. sheet resistivity gives for 2·0 mil width 40 mils length a resistor of 20 K-Ohms while $40 \times 0{\cdot}2 = 8$ sq. mils can give at least eight MOS transistors assuming a fabricational limit of one mil square for each while the MOST, as the load resistor, can have an output resistance easily of the order of 140 K-Ohms or more. Thus a seven times larger resistor with eighth area consumption, i.e., an equivalent of a factor of at least 56 in cost × performance improvement is obtained in MOS circuits on this count in addition to 7 more transistors, that is, for more circuit complexity as well. Also this virtually eliminates the distributed capacitance of the bipolar resistor. Again, MOS transistor has high input impedance greater than 10^{14} Ohms hence no input current is required for fan-outs except for input condenser charging currents. Thus we see that MOS technology has higher possibility of low cost ICs than bipolar technology. Hence the engineering world faces the arsenal to study the possibility of MOS transistor as a load of a driver MOST, their linearity of operation and switching times. On the other hand, since the Bipolar technology is well-developed, one needs to examine the possibility of MOS-Bipolar interfacing which may lead to use of better features of both the types. Since ±5 volts Bipolar ICs are being widely marketed, one needs to examine ways of keeping MOS transistor V_T near ±1 volt. In this pursuit, low power dissipation and high speed is to be necessarily aimed at. Speed, on the other hand, requires parasitic capacitance lowering of interconnection lines and lowering of source drain and gate drain capacitances. Lowering parasitic capacitance requires the interconnection lines to run over thicker oxide layers while the substrate needs to be connected to highest voltage of the system such that the isolation diodes are reverse biased always. This high voltage should be such that it should be nowhere near that producing inversion lines on silicon underneath the interconnection lines over thick oxides. Hence we come across a new parameter the thick field threshold voltage V_{tf} which should have a suitable high value. On the other hand, if a MOST is to work in triode mode, one has to have another supply to put the gates at a higher voltage V_{gg} than the supply voltage V_{dd}. Hence with safety factors one aims at $V_{dd}=3V_T$, $V_{gg}=6V_T$ and a high (V_{tf}/V_T) ratio. This is done keeping in mind that at least $\pm 50\%$ spread can easily occur in the worst cases. Additionally, for the lowering of gate-source and drain-source capacitances, aligning of the masks in step (v) and step (vii), discussed above, are critical. In that, if there is any minute clearance between the gate edge to source diffusion or drain diffusion, then the corresponding capacitance would be small but that there will have a risk of formed channel not connecting either one or both of the source and drain diffusions when the channel is formed under the gate with proper bias on it. Hence, alignment of the source-gate and drain-gate regions just to the required degree of overlap without unduly increasing these capacitances by overdoing is to be aimed at.

Moreover, in this article, we have started on *p*-channel MOST fabrication. The reason being the positive surface charges Q_{sso}, in the early attempts of fabricating *n*-channel enhancement mode MOST led to failures, giving depletion mode MOST since Q_{sso} induced negative mobile charges producing inversion. While depletion mode transistor in the zero gate input state entail power dissipation hence are not suitable for switching circuit applications as are the JFETs. Therefore, it needs to be seen whether surface states can be controlled by technological processes to give the required threshold voltage for enhancement mode operation with *n*-channel FETs. That is we need to study the technological inventories to realise their impact on the control of the above mentioned parameters as well.

11.13 Load Driver Pairs

A MOS transistor can have either a P-channel or an N-channel while the load transistor can be used in either saturation mode or in triode mode over and above it can be an enhancement mode transistor or a depletion mode transistor. Hence one can speak of P-channel and N-channel load driver pairs, as shown in the table below including a special combination of complementary channel MOS transistors both in enhancement mode. The different combinations are :

Designation	Channel	Working mode of load
PELS	P	Enhancement mode load in saturation
PELT	P	Enhancement mode load in triode
PDLS	P	Depletion mode load in saturation
NELS	N	Enhancement mode load in saturation
NELT	N	Enhancement mode load in triode
NDLS	N	Depletion mode load in saturation
CMOS	P and N	Enhancement mode triode to saturation

Case I : PELS

Fig. (11.13.1a) shows the load driver pair connection for PELS & PELT. The load is in saturation when load gate is switched to B and is in triode mode when the load gate is switched to A. For $|V_{gg} - V_T| \leqslant |V_{dd}|$ the load is in saturation and $|V_{gg} - V_T| > |V_{dd}|$ for load in triode (cf. Fig. 11.2.2). Fig. (11.13.1b)

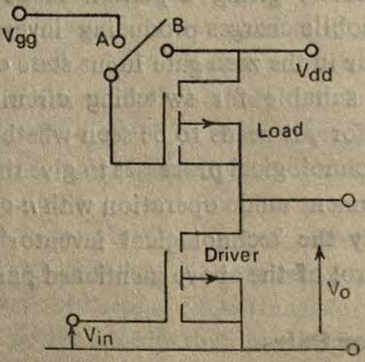
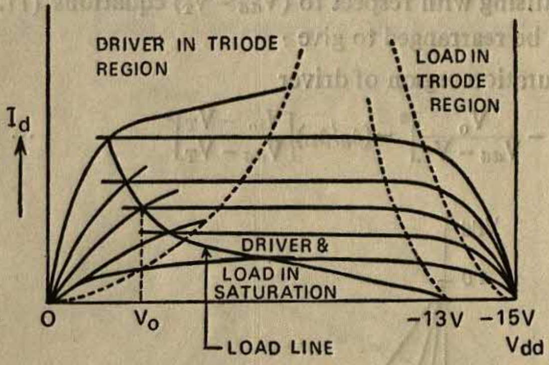


Fig. 11.13.1(a). Load driver pair for PELS (Switch in position B) and PELT (switch in Position A)

shows for a voltage $V_{aa} = -15V$, V_T driver $= V_T$ load $= -2V$ and $(\beta_{driver}/\beta_{load}) = 9$; the superposed output currents for PELS, drawn from left to right for the driver and from V_{aa} to left for the load.



Fig, 11.13.1(*b*). Shows the superposed driver and load output curves intersect to give the load line that the load transistor is acting like.

Any voltage $V_{output} = V_o$ implies it is the voltage across the driver while $(V_{aa} - V_o)$ is the voltage across the load transistor. This figure also shows the non-linear load line for the PELS, the load line starts from a voltage V_T away from the origin of load transistor output characteristic as needed for the gate connected to drain. It is to be noted that in most of the output voltage range the driver is in saturation region.

The theoretical expression for current in the saturated load MOST can be written from equation (11.2.5) as

$$i_L = \frac{\beta_L}{2}\Big[(V_{aa} - V_o) - V_T \Big]^2 \quad \cdots \quad (11.13.1)$$

whereas we have seen that the driver can be in triode or in saturation region while always the driver current must be equal to the load current. When the driver is in saturation

$$|V_{in} - V_T| \leq |V_{out}|$$

and $$\frac{\beta_L}{2}\Big[(V_{aa} - V_o) - V_T \Big]^2 = \frac{\beta_a}{2}(V_{in} - V_T)^2 \quad \ldots \quad (11.13.2)$$

while if the driver is in triode region

$$|V_{in} - V_T| > |V_{out}|$$

and
$$\frac{\beta_L}{2}\Big[(V_{dd}-V_o)-V_T\Big]^2=\frac{\beta_d}{2}\Big[2(V_{in}-V_T)V_o-V_o^{\,2}\Big] \qquad \cdots \quad (11.13.3)$$

Normalising with respect to $(V_{dd}-V_T)$ equations (11.13.2) and (11.13.3) can be rearranged to give :

for saturation region of driver

$$\left[1-\frac{V_o}{V_{dd}-V_T}\right]^2=(\beta_d/\beta_L)\left[\frac{V_{in}-V_T}{V_{dd}-V_T}\right]^2 \qquad \cdots \quad (11.13.4)$$

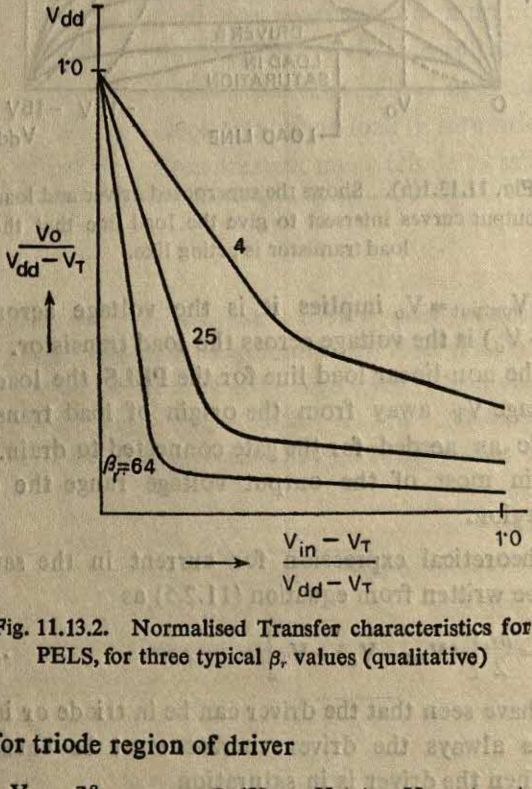


Fig. 11.13.2. Normalised Transfer characteristics for PELS, for three typical β_r values (qualitative)

while for triode region of driver

$$\left[1-\frac{V_o}{V_{dd}-V_T}\right]^2=(\beta_d/\beta_L)\left[2\left(\frac{V_{in}-V_T}{V_{dd}-V_T}\right)\left(\frac{V_o}{V_{dd}-V_T}\right)-\left(\frac{V_o}{V_{dd}-V_T}\right)^2\right] \qquad \cdots \quad (11.13.5)$$

For a given β ratio the PELS pair transfer characteristics that is normalised output $(V_o/\overline{V_{dd}-V_T})$ corresponding to a normalised input $(V_{in}-V_T)/(V_{dd}-V_T)$ can be calculated and plotted as shown in

Fig. (11.13.2) for three typical β ratios. It is to be noted that $\beta_r = (\beta_d/\beta_L)$ the β ratio implies the (width/length) ratio (cf. equation 11.2.6) of the transistors. Higher is the β ratio higher is the gain, i.e., steeper is the slope of the transfer characteristic and lower is the ON voltage of the driver. Consequently, the logical zero level is dependent on the β ratio. This is to be expected from Fig. (11.3.1*b*) since a higher β ratio will give effectively a higher load resistance. In the more or less linear region equation (11.13.4) suggests gain is equal to $-\sqrt{\beta_r}$. If the substrate of the load is connected to ground, as the output increases, the load V_T increases causing the gain to increase but the output swing decreases since it is about $(V_{dd} - V_T)$. This can cause deterioration of output load capacitance charging speed since load allows lesser current to flow due to increase of V_T.

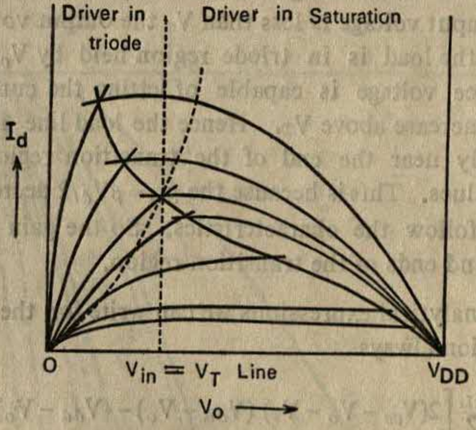


Fig. 11.13.3. Shows the superposed driver and load output curves for MOS load driver pair load in Triode.

Case II : PELT

Next we consider the case of load transistor always in triode i.e., $|V_{gg} - V_T - V_o| > |V_{dd} - V_o|$ which implies $|V_{gg} - V_T| > |V_{dd}|$ i.e., the load gate is connected to a separate supply whose amount of voltage is larger than drain voltage magnitude by more than the value of V_T. This situation corresponds to the gate of load transistor switched to the point A in fig. (11.13.1*a*) when again for a certain

β ratio the superposition of the output characteristics are shown in fig. (11.13.3) which shows the effective non-linear load line also. It is to be noted from fig. (11.13.3) that though the load transistor is in triode always held by large gate supply V_{gg} the driver transistor does work for both saturation in the upper range of output voltage and triode region in the lower ranges of the output voltage. Also note that when the input voltage is less than V_T the output voltage is V_{dd}, this is because the load is in triode region held by V_{gg}, hence very small gate source voltage is capable of letting the current through if V_{in} tries to increase above V_T. Hence the load line is non-linear and particularly near the end of the transition regions acts as of smaller load values. This is because the $g_m = \beta V_d/2$ decreases and the load tends to follow the characteristics. So the gain is lower at the beginning and ends of the transition region.

To find analytical expressions we can write for the load which is in triode region always

$$i_L = \frac{\beta_L}{2}\left[2(V_{gg} - V_o - V_T)(V_{dd} - V_o) - (V_{dd} - V_o)^2\right] \cdots$$

$$= \frac{\beta_L}{2}(V_{dd} - V_o)\,[2(V_{gg} - V_T) - V_{dd} - V_o] \qquad \cdots \quad (11.13.6)$$

with a little rearrangement of terms, equation (11.13. 6) can be written as

$$i_L = \frac{\beta_L}{2}\frac{V_{dd}^2}{m}\left(1 - m\frac{V_o}{V_{dd}}\right)\left(1 - \frac{V_o}{V_{dd}}\right) \qquad \cdots \quad (11.13.7)$$

where $m = [V_{dd}/\{2(V_{gg} - V_T) - V_{dd}\}]$ is the biasing parameter. For triode region of operation of the load $|V_{gg} - V_T| > |V_{dd}|$, when the magnitude of $(V_{gg} - V_T)$ decreases, m increases, becomes equal to unity when $|V_{gg} - V_T| = |V_{dd}|$.

This is rather the boundary between triode and saturation for the load i.e., it is barely in triode. On the other hand, when V_{gg} is very large the other limiting value of m is zero. That is for m between zero to unity the load is in triode region. For typical value of voltages $V_{dd} = 15$ volts, $V_{gg} = 27$ volts and $V_T = 4$ volts $m \sim 0.5$ (cf. ref. 3). If the driver is in saturated mode i.e., $|V_{in} - V_T| \leqslant |V_o|$

$$\frac{\beta_L}{2m}V_{dd}^2\left(1 - \frac{mV_o}{V_{dd}}\right)\left(1 - \frac{V_o}{V_{dd}}\right) = \frac{\beta_d}{2}(V_{in} - V_T)^2 \qquad \cdots \quad (11.13.8)$$

On the other hand, if the driver is in triode regions,

i.e. $|V_{in} - V_T| > |V_o|$

$$\frac{\beta_L}{2m} V_{dd}^2 \left(1 - \frac{mV_o}{V_{dd}}\right)\left(1 - \frac{V_o}{V_{dd}}\right) = \frac{\beta_d}{2}\left[2(V_{in} - V_T)V_o - V_o^2\right] \quad \cdots \; (11.13.9)$$

Equations (11.13.8) and (11.13.9) can be written in normalised form as for saturated driver

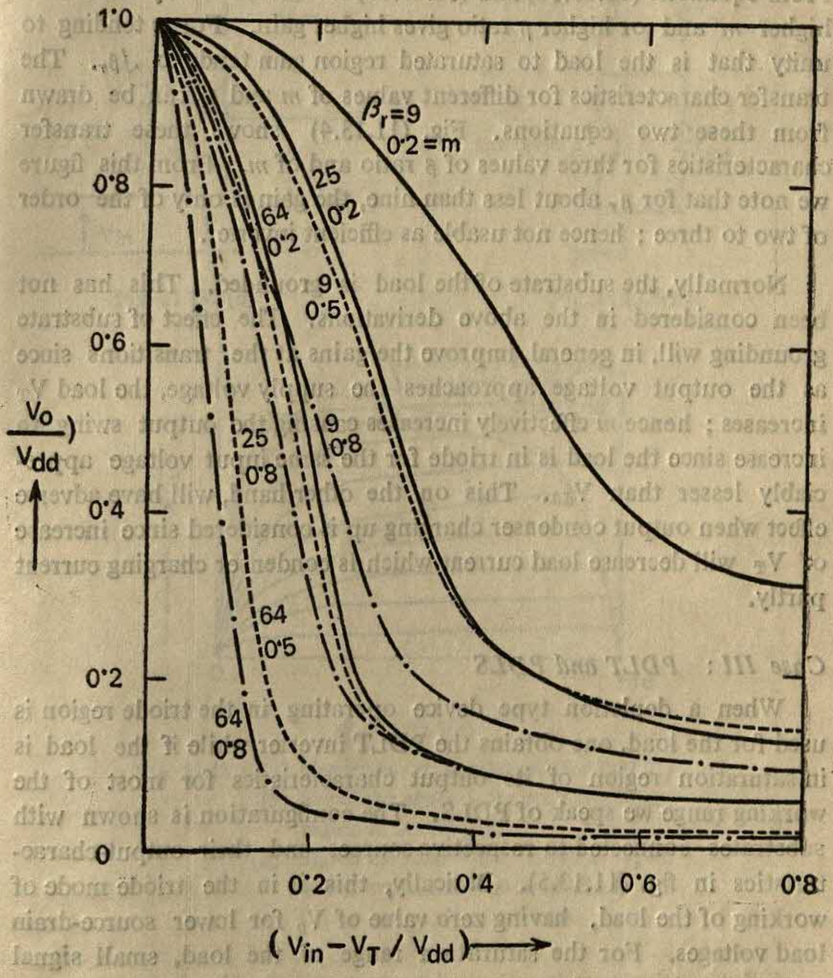


Fig. 11.13.4. Shows the transfer characteristics for values of $\beta_r = 9$, 25 and 64, bold lines for values of $m = 0{\cdot}2$, dotted for $m = 0{\cdot}5$ and chain dot lines for $m = 0{\cdot}8$ in PELT.

$$\left(1-\frac{mV_o}{V_{dd}}\right)\left(1-\frac{V_o}{V_{dd}}\right)=m\beta_r\left(\frac{V_{in}-V_T}{V_{dd}}\right)^2 \quad \cdots \quad (11.13.10)$$

for triode region of driver

$$\left(1-\frac{mV_o}{V_{dd}}\right)\left(1-\frac{V_o}{V_{dd}}\right)=m\beta_r\left[2\frac{(V_{in}-V_T)}{V_{dd}}\cdot\frac{V_o}{V_{dd}}-\left(\frac{V_o}{V_{dd}}\right)^2\right] \quad \cdots \quad (11.13.11)$$

From equations (11.13.10) and (11.13.11) it can be easily seen that higher m and or higher β ratio gives higher gain. For m tending to unity that is the load to saturated region gain tends to $\sqrt{\beta_r}$. The transfer characteristics for different values of m and β_r can be drawn from these two equations. Fig. (11.13.4) shows these transfer characteristics for three values of β ratio and of m. From this figure we note that for β_r about less than nine, the gain is only of the order of two to three ; hence not usable as efficient inverter.

Normally, the substrate of the load is grounded. This has not been considered in the above derivations. The effect of substrate grounding will, in general, improve the gains at the transitions since as the output voltage approaches the supply voltage, the load V_T increases ; hence m effectively increases causing the output swing to increase since the load is in triode for the same input voltage appreciably lesser than V_{dd}. This on the other hand, will have adverse effect when output condenser charging up is considered since increase of V_T will decrease load current which is condenser charging current partly.

Case III : PDLT and PDLS

When a depletion type device operating in the triode region is used for the load, one obtains the PDLT inverter while if the load is in saturation region of its output characteristics for most of the working range we speak of PDLS. The configuration is shown with substrates connected to respective sources and their output-characteristics in fig. (11.13.5). Basically, this is in the triode mode of working of the load, having zero value of V_g for lower source-drain load voltages. For the saturation range of the load, small signal gain expression may be used, when the two generators are as shown in Fig. (11.13.5*b*). In the present ungrounded load substrate, (the $g_2\delta V_T$ generator, which is due to change of V_T of the load if its substrate is grounded is zero), the small signal gain is $\frac{r_{dd}r_{dL}g_1}{r_{dd}+r_{dL}}$. In a depletion mode transistor when back substrate bias increases, it decreases the existing channel strength ; hence V_T decreases. In such case, when the output voltage V_o increases, back gate bias increases thus decreasing V_T, hence current increases through the depletion mode transistor, that is, the effective load value continually decreases due to output voltage rise. This straightens up the load line a bit as

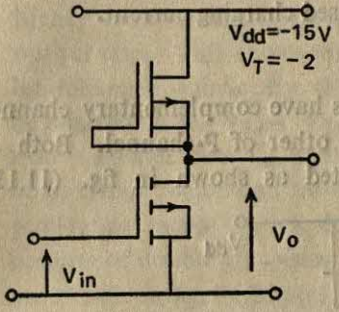


Fig. 11.13.5(*a*)

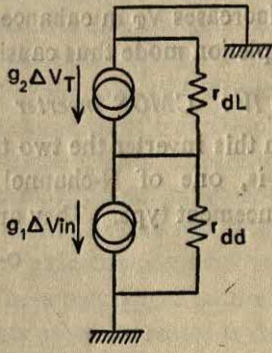


Fig. 11.13.5(*b*)

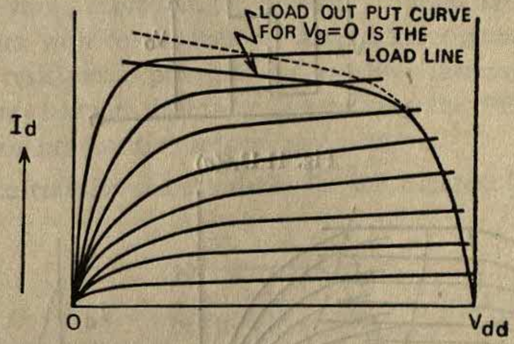


Fig. 11.13.5(*c*)

Fig. 11.13.5. (*a*) Shows the driver load pair for PDLT ;
(*b*) Shows the equivalent circuit for linear region of operation ;
(*c*) Shows the load output curve of $V_g=0$ which represents the load line to the driver output characteristics.

shown by the dotted line. A value of $\beta_r \sim 30$ is about the maximum (cf. ref. 4) for positioning the operation in the triode region which, of course, will depend on bias. The large signal gain depends on β_r. Value of $\beta_r \sim 20$ (cf. ref. 4) is enough to give high charging rates for output condenser loads. The load condenser charging is expected to be better in this case than the previous cases since back substrate bias increases V_T in enhancement mode transistors while it decreases in depletion mode thus causing increased charging current.

Case IV: CMOS Inverter

In this inverter the two transistors have complementary channels, that is, one of N-channel and the other of P-channel. Both are enhancement type. They are connected as shown in fig. (11.13.6)

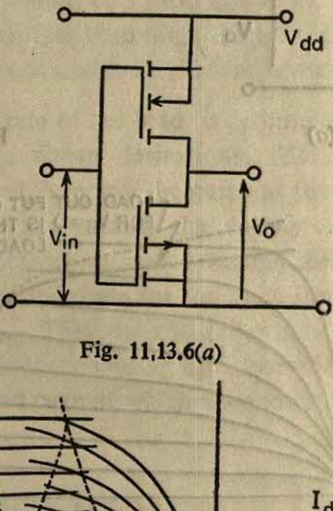


Fig. 11.13.6(*a*)

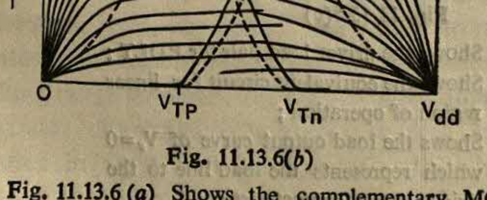


Fig. 11.13.6(*b*)

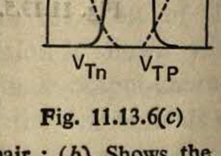


Fig. 11.13.6(*c*)

Fig. 11.13.6 (*a*) Shows the complementary MOS pair ; (*b*) Shows the output characteristics with their triode $V_g - V_T > V_d$ and saturation $V_g = V_d$ dotted lines as also the current pulse shape solid line when $|V_{TP}| + |V_{Tn}| < |V_{dd}|$; (*c*) Shows the saturation $V_g = V_{dd}$ dotted lines and current (solid curve) when $|V_{Tn}| + |V_{TP}| > |V_{dd}|$.

with their output characteristics and load lines. Fig. 11.13.6(*a*) shows that their gates are connected to each other. When V_{in} is zero, the driver is in saturation but the upper transistor is in triode ; thus for output voltage rise the lower transistor presents a high resistance load to the upper one while when V_{in} is V_{dd} the lower transistor is in triode while the upper one is in saturation. Thus, when output voltage falls, the upper transistor increasingly presents a higher load to the lower driver. While the opposite is the case on output rise. This shows that the roll of the load and driver can be interchanged. Obviously, the best gain and fastest transition would be be obtained when $V_{dd} = |V_{TN}| + |V_{TP}|$ since if this threshold voltage sum is greater than V_{dd} then for the middle output range both would be in triode region and the behaviour would be like PELT in low output range except for a doubling of gain for identical transistors because of double transistor drive. These overlapped saturation lines are shown in fig. 11.13.6(*c*) ; larger this overlap greater is the current pulse size and duration in the transistor. For correct supply voltage when one transistor enters in saturation, the other enters in triode and vice versa, hence transfer linearity is preserved. For identical transistors with correct voltages larger current generator with saturated load resistance is presented for transition increasing the gain by a factor, may be more than twice larger, over the PELS while power dissipation in transition reduces enormously.

The transfer characteristics for the different load driver pairs

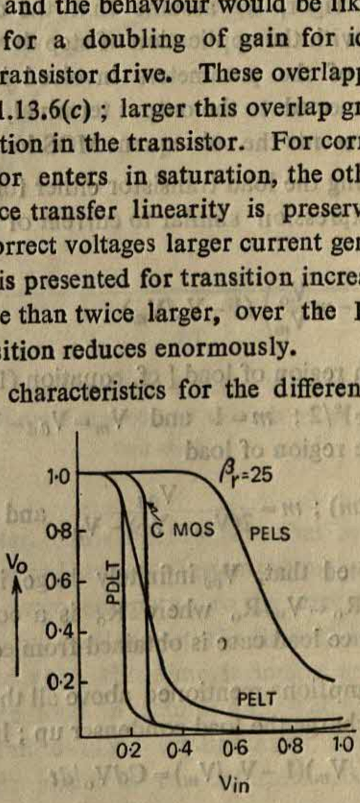


Fig. 11.13.7 Shows the fractional logic output swing vrs. fractional logic input swing for different driver load pairs for $\beta_r=25$.

are compared in Fig. (11.13.7) for normalised output swing which shows that PDLT and CMOS have comparable gains while PELT has lower and PELS has still lower gain. Proper matched CMOS gain will be independent of β_r ; hence, it is called ratio less inverter. It is costly since fabrication of identical V_T complementary channel on same substrate requires well controlled MOS technology or other costly technology to be discussed later.

Rise time consideration

For an understanding of the operation of the load driver pairs in switching application, we consider here only one part of the switching phase, that is, in the rise of the output voltage. For simplicity avoiding rigor, we assume that the driver transistor is off instantaneously with the application of step drive input V_{in}. This is only to get a feel of the parameters involved in charging a condenser which is connected to the output terminal, the condenser, on the other hand, represents the subsequent MOS loads.

For representing the load transistor either for saturation or for triode region, an expression similar to current of equation (11.13.7) will be used as

$$i_L = I\left(1 - m\frac{V_o}{V_m}\right)(1 - V_o/V_m) \quad \cdots \quad (11.13.12)$$

where for saturation region of load [cf. equation (11.13.1)]

$I = \beta_L(V_{dd} - V_T)^2/2$; $m = 1$ and $V_m = V_{dd} - V_T$ the maximum swing and for triode region of load

$$I = (\beta_L V_{dd}^2/2m) \; ; \; m = \frac{V_{dd}}{2(V_{gg} - V_T) - V_{dd}} \text{ and } V_m = V_{dd}.$$

It is to be noted that, V_{gg} infinitely large implies $m = 0$ and that ; with $I = V_{dd}/R_o = V_m/R_o$ where R_o is a constant resistance, the constant resistance load case is obtained from equation (11.13.12).

With the assumptions mentioned above all the current supplied by the load goes to charge the load condenser up ; hence

$$I(1 - mV_o/V_m)(1 - V_o/V_m) = C\,dV_o/dt \qquad (11.13.13)$$

dividing both sides by V_m and using $V_m/I = R$, (which is as if the short circuit resistance of the load from its maximum voltage at output level to ground), we have on transposing

$$\frac{dt}{\mathrm{T}}=\frac{dv_o}{(1-mv_o)(1-v_o)} \quad \cdots \quad (11.13.14)$$

where substitution of $v_o=(\mathrm{V_o}/\mathrm{V}_m)$ and $\mathrm{T}=\mathrm{CR}$ has been made. On integration between $t=0-$ to t we have from (11.13.14) after rearrangement

$$\frac{\mathrm{V_o}}{\mathrm{V}_m}=[1-\exp-(1-m)t/\mathrm{T}]/[1-m\exp-(1-m)t/\mathrm{T}]$$

$$\cdots \quad (11.13.15)$$

equation (11.13.15) is plotted as a function of (t/T) in Fig. (11.13.8)

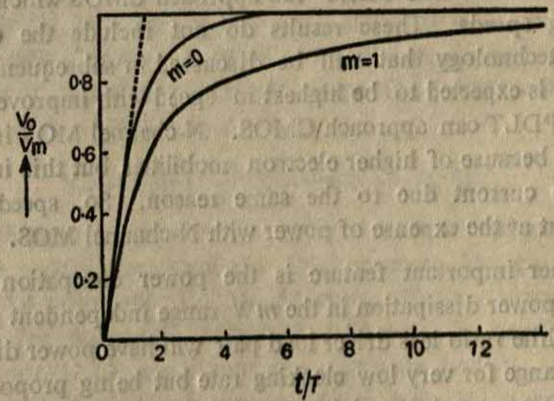


Fig. 11.13.8. Idealised case condenser load charging logic rise times for resistive $m=0$ load through triode to saturation $m=1$.

with m as a parameter. This figure shows that even in ideal case of the driver current stopped to zero instantaneously by drive (say), the output rise time is large compared to the case of $m=0$ or initial resistance presented by the load at the time of switching the input voltage down. Also that, this time is longer as m tends to unity corresponding to the load in saturation. This is obviously to be expected since saturation mode implies larger load resistance than load in triode mode and that the saturated load limits charging current supply.

Detailed computer modelling of transient performance of driver load pairs by Carr and Mize leads them to conclude (cf. ref. 4).

That the load in ratio type pair has a reduced g_m because of its lesser β ; hence slower in charging the output condenser. The PELT load being in triode is fastest of the ratio types in condenser charging phase. PELS is slower since the load is in saturation and saturation resistance is larger than triode resistance.

On the other hand, PELT and PELS fall times are much faster than their rise times since the driver goes deep in triode in both cases while PELS load output resistance being large allows condenser discharge faster. The PDLT and CMOS inverters with proper selection of g_m and V_{th}, i.e., for proper design can give nearly equal rise and fall times while PDLT can approach CMOS which is fastest in charging speeds. These results do not include the effects of improved technology that will be discussed in subsequent articles. The CMOS is expected to be highest in speed with improved technology while PDLT can approach CMOS. N-channel MOS is expected to be faster because of higher electron mobilities but this is expected to increase current due to the same reason. So, speed may be increased but at the expense of power with N-channel MOS.

Another important feature is the power dissipation ; PELT, PELS have power dissipation in the *m*W range independent of clocking speed while ratio less driver load pair will have power dissipation in the *n*W range for very low clocking rate but being proportional to clocking rate approaches and becomes about one *m*W around 1 MH*z* and 10 *m*W at 10 MH*z*. While the normal RTL and TTL is constant around 10 *m*W upto about 50 MH*z*, the former being higher and the latter lower in this range. ECL, on the other hand, remains about constant upto about few tens of MH*z* around 20—30 *m*W. For further details the reader is referred to the original publication of the above-mentioned authors.

11.14. MOS Technology

In this article we consider the different technological methods and their impact on the control of MOS threshold voltage, gate source and gate drain capacitances ; hence device speeds.

Firstly, to get an idea of these parameters as obtainable with the standard (111) process, that is discussed in article (11.12), we start with the achievable physical and dimensional parameters with standard bipolar technology at room temperature T = 300°K. The value of $\varnothing_{MOX}$ for Al-SiO$_2$ system was stated to be 3·2 volts and $\varnothing_{SOX} = 3{\cdot}25$ for Si-SiO$_2$ system in article (11.6) using an N-substrate of doping level of $2{\cdot}9 \times 10^{15}$/cc. We have from equation (11.5.1) $\varnothing_f = -0{\cdot}29$ volts with $n_i = 2{\cdot}5 \times 10^{10}$/cc and $kT/q = 0{\cdot}025$ volts. Assume an oxide thickness of $t = 1000°$A its dielectric constant 4, when we have $C_{ox} = 3{\cdot}54 \times 10^{-8}$ F/cm^2. A surface charge of value $Q_{ss} = 4{\cdot}4 \times 10^{11}.q/\text{cm}^2$ is typical in (111) orientation of silicon (q being the electronic charge $= 1{\cdot}6 \times 10^{-19}$ coulomb).

From equation (11.6.2), we have $\varnothing_{MS} = 3{\cdot}2 - (3{\cdot}25 + 0{\cdot}55 - 0{\cdot}29) = -0{\cdot}31$ volt. From equation (11.6.13) $V_{fb} = -0{\cdot}31 - (+2) = -2{\cdot}3$ volts. While $Q_{sdm} = \sqrt{[4\varepsilon_0 \varepsilon_s q N_d |\varnothing_f|]} = 2{\cdot}22 \times 10^{-8}$ coul/cm^2, hence $V_T = -2{\cdot}3 - 0{\cdot}58 - 0{\cdot}627 = -3{\cdot}507 \simeq -3{\cdot}51$V. Assuming a thick oxide layer of 15000°A, thick field oxide $C_{ox} = 0{\cdot}236 \times 10^{-8}/\text{cm}^2$; $V_{fb} = -0{\cdot}31 - 30 = -30{\cdot}31$V and

$$V_{tf} = -30{\cdot}31 - 0{\cdot}58 - 9{\cdot}405 = -40{\cdot}3\text{V}.$$

Hence, including tolerance on Q_{ss} and production tolerance totalling as much as ±25% the standard (111) process V_T ranges from $-2{\cdot}5$ to $-4{\cdot}4$ volts and V_{tf} 30—50 volts. Hence, power supply requirement for N-channel MOS/LSI fabricated by standard (111) process (cf. art. 11.12) are about $V_{dd} = 12$ volts and $V_{gg} = 22$ volts. The higher value of V_T is coming from high V_{fb} that is from (Q_{ss}/C_{ox}).

Next, consider another crystal orientation, i.e., (100) plane to lower V_{fb}. On this orientation, the value of Q_{ss} is lower than in (111) plane by about a factor of 6.

When $V_{fb} = -0{\cdot}31 - 0{\cdot}33 = -0{\cdot}64$, while $V_T = -0{\cdot}64 - 0{\cdot}58 - 0{\cdot}627 = -1{\cdot}85$. For thick oxide region $V_{fb} = -0{\cdot}31 - 5{\cdot}0 = -5{\cdot}3$V, $V_{tf} = -5{\cdot}3 - 0{\cdot}58 - 9{\cdot}405 = -15{\cdot}29$ with ±25% tolerance, $V_T = 1{\cdot}4$ to 2·3V, $V_{tf} = 11$ to 19V, while supply voltage requirement is $V_{dd} = -6$V, $V_{gg} = -12$V.

The advantages are that the low threshold voltage is suitable. Also, lower power dissipation possibility is there because of lower supply voltage V_{dd}. While, if the thick field V_{tf} was a little higher, it would have been better to avoid V_{gg} voltage fluctuations not crossing the V_{tf} limits. Moreover, the mobility on the (100) plane is lower by about a factor of 0·80 which means, to keep the circuit speed to the same level requires lowering the channel length to about 90% (cf. equation 11.2.10).

Another approach to lower V_{fb} in an effort to find bipolar compatibility is increasing C_{ox}. Silicon nitride (Si_3N_4) has higher relative permittivity of 7·5 as opposed to 4 of SiO_2. Hence on (111) plane of silicon on top of about 200°A SiO_2 (which thus preserves the Q_{ss} value to the same $4{\cdot}4 \times 10^{11} q/\text{cm}^2$), an 800°A Si_3N_4 is grown then, effectively a dielectric constant of 6·8 results for the same total oxide thickness, which thus results in increase of C_{ox} by a factor of about 1·7. Chemical vapour deposition techniques are well developed where reaction between silane and ammonia produce uniform nitride films of good quality accoding to $Si\,H_4\,(gas) + 4N\,H_3\,(gas) \rightarrow Si_3N_4 + 12H_2\,(gas)$. The reaction temperature ranges between 800—900°C.

Direct growth of nitride on silicon gives carrier trapping at the semiconductor insulator surface that causes shift of V_T with gate bias. Hence, a layer of SiO_2 of the order of 200°A is used first, which decreases this effect. This is known as the 'Nitride Process'.

In this $V_{fb} = -0{\cdot}31 - 1{\cdot}2 = -1{\cdot}51V$, $V_T = -1{\cdot}51 - 0{\cdot}58 - 0{\cdot}37 = -2{\cdot}46\ V$, $V_{fb} = -0{\cdot}31 - 24 = -24{\cdot}31$ (since from $\frac{1}{c} = \frac{1}{c_1} + \frac{1}{c_2}$, ε_r = about 5 results), $V_{tf} = -24{\cdot}31 - 0{\cdot}58 - 7{\cdot}52 = -32{\cdot}41$.

Hence the spread of values of

$$V_T - 1{\cdot}9 - 3V,\ V_{tf} - 25 - 40,\ V_{dd} = -7{\cdot}5V,\ V_{gg} = -15V$$

In this technology, we see high V_{tf} and possibility of interfacing with bipolar, high (111) surface mobility, high ε_r ; hence high g_m are all obtained but since number of process steps is increased the yield decreases ; hence cost increase is evident.

Another technology known as the 'Polysilicon Gate Process' resulted from the effort of controlling V_T by changing ϕ_{MOX}. In this, instead of Al-metallisation, polycrystalline silicon is deposited on the silicon-dioxide and then P-doping is made for the source drain regions after defining, exposing and etching these regions when simultaneously the polysilicon gets doped heavily with P-type impurity giving very good conductivity with sheet resistances less than 100 ohm per square. This has two facets of control. Firstly, the ϕ_{MOX} of heavy doped polysilicon with respect to SiO_2 is about 4·2. Thus V_T is directly affected by about plus one volt since ϕ_{MS} becomes now

+0·7 volt. Thus V_T becomes −2·52 volts while $V_{tf} = -39{\cdot}3$ volts when range of V_T becomes −1·9 to −3·1, while V_{tf} ranges from −29 to −50V, $V_{aa} = -7{\cdot}5$ volts and $V_{gg} = -15$ volts. These are quite satisfactory from the point of view of the interfacing possibility with low power dissipation. The second great advantage of this process is that doping of the source drain is being done after etching out windows through polysilicon and the underlying SiO_2. This is possible because of polysilicon's refractory properties. Al would be evaporated when exposed to temperatures required for doping (i.e., of the order and above 1000—1250°C). While polycrystalline silicon would stand this temperature safely as a mask additionally, this doped polysilicon can be used as the gate metal.

The direct consequence of this source drain diffusion is that the polysilicon works as the mask where overlap of source drain regions with the gate edge on silicon surface is self controlled and gives a smaller overlap. The gate source overlap being limited to lateral diffusion and gives overlap of the order of 1·7 microns. Reduction of overlap capacitance improves the circuit speed by about 50% over that would have otherwise resulted in a separate aligning step (this step was critical too). Moreover, it becomes easy to incorporate double layer interconnection by oxidising polysilicon over the portions where interconnection lines are laid and dividing this oxide in two portions as the source and the drain sides. Since the number of process steps decreases the economy improves even if this process is more involved.

Another important method is positive ion implantation within the crystal lattice by using an ion accelerator with its focussing and scanning systems. In this method ionised dopant atoms are accelerated by voltages of the order of 50—200 kv and allowed to impinge on the crystal surface. The ions enter the lattice, the depth of penetration being dependent on the accelerating potential. The bombardment of the crystal with ions causes surface damages which can be cured by proper temperature annealing in the range of 400—500°C for about 30—40 minutes. The doping profile by ion implantation can be well-controlled by varying the accelerating potential and the beam current.

The distribution of implanted ions on silicon surface region is approximately Gaussian and is given by

$$N(x) = \frac{D}{\delta R_p \sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x - R_p}{\delta R_p}\right)^2\right]$$

where R_p is called the projected range, so called because this is not necessarily the penetration depth for normal incidence on the surface, the straggle δR_p, on the other hand, is a measure of the spread of the distribution and the dose is $D = \int_0^\infty N(x)\,dx$. The concentration distribution depends on the temperature of the silicon substrate as also on the atomic mass ratio of substrate atom and implanted atom. Greater this ratio is less is $(\delta R_p/R_p)$ the relative straggle. From this it is evident that the range and straggle for ions like boron, phosphorus or Arsenic will be largely dependent on accelerating potential (cf. ref. 4) as also crystal orientation. Nevertheless, when proper alignment has been made, the above distribution can be obtained for random implants. Hence, one can choose the accelerating potential such that the ions can penetrate through oxide thickness of the order of 1000°—1500°A and can get implanted in underlying silicon substrate but fails to penetrate thick oxide regions or thin film metallic layers more than a few hundred angstorms obtained by vacuum evaporation. Moreover, the ion beam current can be monitored from which the total number D/cm^2 can be accurately calculated. Thus, after growth of the thin oxide, the surface doping under the thin oxide regions can be done to the required level to even within 1% and after placing the metal gate on the thin oxide source drain doping can be obtained. Thus surface V_T control alongwith perfect alignment is obtained with ion implantation technique.

Assuming that the straggle through oxide implant is within 0·15 micron of the silicon surface for boron ions and a particular accelerating potential, an ion beam with a current density of 2·4 *n*A for ten seconds can implant $1{\cdot}5 \times 10^{+11}$ atoms/cm^2. Assuming uniform doping this would give 10^{16} atoms/cc concentration around the surface region of silicon. This would convert the silicon surface to P-type with a dopant concentration of $7{\cdot}5 \times 10^{15}$/cc. This would give a $\varnothing_f = +0{\cdot}31$ volts, $\varnothing_{MS} = 3{\cdot}2 - (3{\cdot}25 + 0{\cdot}55 + 0{\cdot}31) = 3{\cdot}2 - 4{\cdot}11 = -0{\cdot}91V$, $V_{fb} = -0{\cdot}91 - (+2) = -2{\cdot}91V$

while $Q_{sdm} = \sqrt{[4\varepsilon_0 \varepsilon_s q N_a \phi_f]} = 3{\cdot}97 \times 10^{-8}$ coul/cm^2,
$V_T = V_{fb} + 2\phi_f + 1{\cdot}122 = -2{\cdot}91 + 0{\cdot}62 + 1{\cdot}122 = -1{\cdot}16$ V.

For the same thick oxide layer thus, $V_{fb} = -0{\cdot}91 - 30 = -30{\cdot}91$V,
$V_{tf} = -30{\cdot}91 + 0{\cdot}62 + 16{\cdot}83 = -13{\cdot}46$, $V_T = 0{\cdot}8 - 1{\cdot}5$, $V_{dd} = -4$V
$V_{tf} = 10 - 17$, $V_{gg} = -7$V

The above example clearly shows that by proper choice of ion beam current or implant time the threshold voltage can be controlled and instead of an enhancement mode device, one can obtain a depletion mode device as well. That is V_T can be made positive by increasing the implant dose. (The student can repeat the computation for an implant current 9·6 *n*A for 30 secs).

Apart from the alignment accuracy and threshold control, ion implantation gives the possibility of fabricating N- and P-MOS simultaneously on the same substrate. That is complementary MOS transistor having matched V_T is obtainable with ion implantation. Enhancement mode driver depletion mode load inverter having required V_T values can be obtained on the same chip. Thus, TTL/MOS interfacing with low power dissipation, and high speed capability inverters can be obtained with implantation technology. The disadvantage is the high initial cost of the sophisticated ion implanter for batch fabrication of devices.

Another important technology needs to be mentioned in this connection is the silicon on saphire (SOS) technology. In this epitaxial growth of silicon on spinel structure particularly saphire is obtained by pyrolysis of silane. This gives single crystalline silicon of 1 or 2 microns on a few miles thick saphire substrate. Saphire is an insulator, its thermal expansion coefficient in certain orientations match that of silicon and favours epi-layer growth of silicon. This epitaxial layer can be utilised for fabrication of N- or P-MOS or both by standard techniques or by ion implantation. Thus either CMOS or PDLT inverters can be obtained.

The basic reason for using insulator as the substrate is that the ultimate speed of devices and circuits are almost always determined by the stray capacitances from the active regions to the snbstrate. In bipolar technology the isolation islands and active regions are coupled through reverse biased junction capacitances to ground. In MOS on silicon substrate the interconnection lines are mutually coupled through capacitances between themselves as also through individual capacitances of these lines to substrate which, therefore, is chosen of high resistivity to minimise the last coupling factor. It is this coupling factor which is enormously reduced if saphire substrate is used which thus reduces the source to drain capacitance that reduces speed due to miller effect capacitance of source and drain to substrate as also the distributed capacitances with interconnection lines through substrate are enormously reduced ; thus, the speed capability of SOS-MOST circuits increases and hence with SOS-CMOS the speed power product that is (switching time × power) product lowers.

Lastly, the P-substrate N-channel devices were left out in our discussions so far. The reason was that Q_{ss} charge is always positive hence gives an inverted channel ; so, a depletion mode transistor always resulted while we have seen in (100) plane the value of Q_{ss} is comparatively lower. Hence, the question arises whether in (100) plane with the sophisticated technology described above, it may be possible to have enhancement mode transistor. To examine these, we firstly determine the threshold voltage V_T for the (100) plane for the same oxide thickness and doping level but of a P-type substrate ; hence $\phi_f = +0.29$ when we have $\phi_{MS} = 3{\cdot}2 - (3{\cdot}25 + 0{\cdot}55 + 0{\cdot}29) = -0{\cdot}89$V and $V_{fb} = -0{\cdot}89 - (+0{\cdot}33) = -1{\cdot}22$V hence $V_T = -1{\cdot}22 + 0{\cdot}58 + 0{\cdot}627 = -0{\cdot}013$V. That V_T is small and negative implies even on (100) plane fabricated MOS a small negative voltage has to be applied for current to be zero ; hence, a depletion mode transistor has resulted ; backgate biasing of this MOST can give enhancement mode of operation easily.

If, instead of Al-gate, polysilicon gate is used, what would be the result ? Remembering that ϕ_{MOX} of polysilicon gate with SiO_2 is 4·2, we have $\phi_{MS} = 4{\cdot}2 - (3{\cdot}25 + 0{\cdot}55 + 0{\cdot}29) = 0{\cdot}11$V, when $V_{fb} = +0{\cdot}11 - (+0{\cdot}33) = -0{\cdot}22$V and hence $V_T = -0{\cdot}22 + 0{\cdot}58 + 0{\cdot}627 = +0{\cdot}99$V.

Hence it is quite possible to have N-channel enhancement mode transistor with appropriately low V_T of the order of a volt by using sophisticated technology. But the next question arises what about the speed of these devices ? The mobility of electrons is twice the mobility of holes, so a device of this sort is expected to have double the switching speed and frequency performance should be better. In practice, however, N-channel MOS did not show a doubling of speed, one possible reason may be partly due to lower mobility on the (100) plane compared to the (111) planes. Increased electron carrier mobility, on the other hand, increases the current whence power dissipation by a factor of two. Thus the speed power product increased to twice that of P-channel devices. Of course, these will give better high frequency performance. Is it then worth having enhancement mode devices using the sophisticated technology mentioned above ?

Depletion mode transistor as a load has obvious advantages,

(1) that there is no need for separate gate voltage supply V_{gg} ;

(2) hence the circuit is simpler also it has zero gate voltage always ;

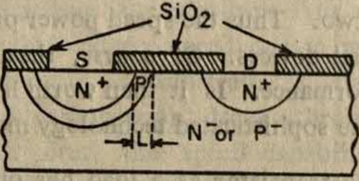
(3) compared to the enhancement mode, the load transistor can have lesser conductivity ; therefore, for bringing the output more near V_{dd} in the output 1 state, since as soon as the driver source drain voltage V_{ds} exceeds V_T it goes in saturation and becomes a constant current generator ;

(4) so the driver needs small channel width W henee has smaller capacitances rsulting in higher speed with the possibility of increased packing density ;

(5) additionally, the depletion mode load is effectively a constant current generator for most of the voltage range of swing for example, in the output rising phase upto $|0-V_T| < |V_{ds}|$ hence it provides the same charging current to the load capacitance even when the voltage across the load transistor is small, this increases the switching speed since almost all of the voltage range is swept with large current ;

(6) the requirement of lesser conductivity points to the usability of P-channel load while a greater packing density with increased switching speed gives two to ten times smaller speed power product. Typically, the value comes down to 10 PJ.

When discussing about speed it is worth mentioning, a few typical structures and techniques that have evolved to increase the switching speed. It has been noted (cf. equation 11.2.10) that the speed or frequency performance of a MOST improves with decrease of chaanel length. The attempt to decrease this length resulted in the D-MOS and V-MOS technologies. In D-MOS technology (the 'D' stands for double diffused) successively two diffusions are made through the same mask opening (cf. fig. 11.14.1). The first diffusion produces a P-type region around the source. Before the second diffusion, the drain diffusion window is also opend and N^+ source and drain diffusions are made when one obtains the channel length L determined by the successive P and N^+ diffusion boundaries. The rest of the region between source and drain under the gate oxide acts only as drift space having N-type high conductivity. This is because the substrate is either lightly doped N or P. Hence the presence of positive surface charges Q_{ss} creates and N-channel in this wider region even without bias. The bias that is needed to invert the channel region (L) augments the channel strength in the rest of the region. Hence the charge relaxation time in this region becomes short enough to exhibit the length L as the effective channel length. Normally the P-region doping is at least one order of magnitude higher than that of the substrate hence of the wider region. This in turn, makes the source drain breakdown voltage high. With this type of MOS, speed power product of the order of 0·25 PJ is attainable.



Fig, 11.14.1. Shows the double diffused MOS. P and N^+ diffusions are successively made through the same source opening to obtain the channel length L small.

The V-MOS, on the other hand, derives its name from the V-shaped anisotropic etch cut on the (100) plane leaving the V-edges on the (111) plane. Firstly, one starts with (100) plane of an N^+ substrate, grows an one micron epitaxial P region followed by one micon P^- region and then an N^+ region. N^+ regions are diffused locally into the surface to form the drains of the transistors. A

V-shaped groove is etched into the silicon surface down through the top N^+ and P layers such that the V edge enters the N^+ substrate below as shown in fig. (11.14.2). Gate oxide is grown on the V-groove walls and gate metal is deposited on it. The N^+ substrate forms the common source, hence this gives extended source surface area over V-apex. The channel length L is obtained along the V body in the P region. The P^- region being inverted due to the presence of positive surface charge Q_{ss}. This region reduces the drain source capacitance and increases the punch through voltage by extending the depletion region from drain to source. The channel width is comparatively large because it encircles the V-groove. Hence with smaller V-opening larger channel width; hence current is obtained. Thus, speed power product of the order of 3 pJ results. Note that the gains for V-MOS transistor having short channel lengths and working in the velocity saturation region is very good because of higher g_m obtained due to higher W, the channel width.

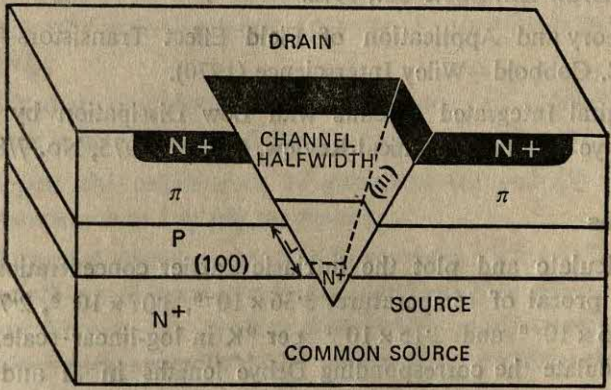


Fig. 11.14.2 Shows V-MOST cross section. The (111) plane specifies the groove. SiO_2 grown on (111) planes within the groove, on which gate metal is deposited. Channel length L and channel half width are shown.

References

1. Physics of Semiconductor Devices by S.M. Sze, Wiley Eastern Ltd., 1979.
2. MOS Field Effect Transistors and Integrated Circuits by Paul Richman, John Wiley & Sons (1973).

3. MOS Integrated Circuits by W.M. Penney & L. Lau, (Van Nostrand Reinhold Company) 1972.
4. MOS/LSI Design and Applications by W.N. Carr & J.P. Mize, McGraw-Hill Book Co., 1972.
5. Theory and Application of Field Effect Transistors—Richard S.C. Cobbold—Wiley Interscience (1970).
6. Digital Integrated Circuits with Low Dissipation by N.C. de Troye—Philips Technical Review, Vol. 35, 1975, No. 7/8, p. 212.

Problems

1. Calculate and plot the intrinsic carrier concentration in Si at reciprocal of temperatures $3{\cdot}36 \times 10^{-3}$, $3{\cdot}07 \times 10^{-3}$, $2{\cdot}75 \times 10^{-3}$, $2{\cdot}45 \times 10^{-3}$ and $2{\cdot}15 \times 10^{-3}$ per °K in log-linear scale. Hence calculate the corresponding Debye lengths in Si and plot in similar scale.
2. Calculate and check the surface depletion region width x_d assuming the surface as intrinsic for bulk doping levels of 10^{14}, 10^{15}, 10^{16} and 10^{17} in Si as 1·8, 0·6, 0·2 and 0·07 microns. Compare these with the maximum depletion widths possible as 2·05, 0·85, 0·3 and 0·1.
3. Calculate the maximum surface depletion region charge density and compare it with the surface depletion region charge density when the surface is intrinsic for the doping levels of problem 2 above in Si. Note that these are given by (Q_{sd}/q) per cm^2 as for maximum 2·5, 8·5, 30 and 100 and for intrinsic surface 1·75, 6. 20 and 70 multiplied by 10^{10} respectively. Plot in log-log scale.
4. Plot in linear-log scale the vatiation of E_F in *e*V and ϕ_F in volts for P-doped and N-doped semiconductors at 300°K for doping levels 10^{11}, 10^{13}, 10^{15} and 10^{17}. Note that the ϕ_F values are approximately +0·05, +0·17, +0·29, +0·41 volts respectively for P-doped and negative but have the same values for N-doped material.
5. Find the metal semiconductor work function difference ϕ_{MS} as a function of doping concentration in Si substrate for Aluminium gate material. Assume $\phi_{MOX} = 3{\cdot}02$ and $\phi_{SOX} = 3{\cdot}25$ *e*V. Note that for P-type substrate the values of ϕ_{MS} corresponding to the

doping levels of problems (4) above are $-0{\cdot}64$, $-0{\cdot}76$, $-0{\cdot}89$ and $-1{\cdot}015$ respectively ; hence plot for the N-type substrate.

6. Find the values of V_T for Al-gate P-substrate MOSFET having $t_{ox} = 1000°A$ at $300°K$ assuming the $(Q_{ss}/q) = 5 \times 10^{11}$ and 10^{12} per cm^2 for doping levels 10^{11}, 10^{13}, 10^{15} and 10^{17} respectively.
7. Calcultate the thick-field threshold voltage $t_{ox} = 12000°A$ for the same values of Q_{SS} and doping levels of problem (6) above.
8. Repeat the calculations of problems (6) and (7) above for N-substrate and justify the results.
9. N-channel depletion MOSFETs are normally made on a high resistivity substrate to minimise back gate bias effect. How can you fabricate an N-channel depletion mode transistor having a V_T widely controllable by substrate to source reverse bias ? Find some numerical values to substantiate your arguments.

Chapter 12

TRANSIT TIME DEVICES

The most important transit time device is Impatt. Impatt refers to Impact Ionisation Avalanche Transit Time devices. In these devices reverse bias field induced impact produces avalanche ionisation in the junction of a diode. When it is mounted in a MW cavity the a.c. voltage produces increased avalanche produced carriers at the positive half cycle. This causes a large number of carrier production over half a cycle. These carriers then propagate through the medium and reach the electrode after a certain transit time. By then the phase of the voltage across the diode changes and if this phase change is of the order of 180° the peak fundamental current is collected around peak negative voltage across the device. That is to the terminal voltage the device presents a negative resistance. If the device is mounted in a suitable MW circuit, the terminal MW voltage can be made self-sustaining.

Such negative resistance arising from transit time in semiconductor diodes was first proposed by Shockley (*ref. 1*). In 1958 Read (*ref. 2*) dealt with a high frequency semiconductor diode in which he considered a reverse biased diode of the form P^+-N-I-N^+ or N^+-P-I-P. In these structures he considered a very narrow avalanche region that was followed by a high resistance region for the drift of avalanche produced carriers. This drift region provided the required transit time for the generated carriers.

The physics of the Read diode operation can be summarised with reference to fig. (12.0.1). If this diode is reverse biased to breakdown and an a.c. voltage is added to the d.c. voltage, the peak field at the P-N-junction increases resulting in avalanche multiplication to increase during the positive half of the R.F. cycle. Later, in the cycle the decrease of the terminal voltage causes the excess multiplication to stop. The carriers generated by this process separate under the influence of the high field, with the holes moving into the P^+ region while the electrons drifting with saturated velocity through the I-region. The drifting electron current induces in the external circuit the constant current shown in fig. 12.0.1(e). From figs. 12.0.1 (c and e) it is clear that the diode exhibits a negative conductance in parallel with a capacitance at its terminals and that an electron transit time half of the period of the a.c. voltage wave-

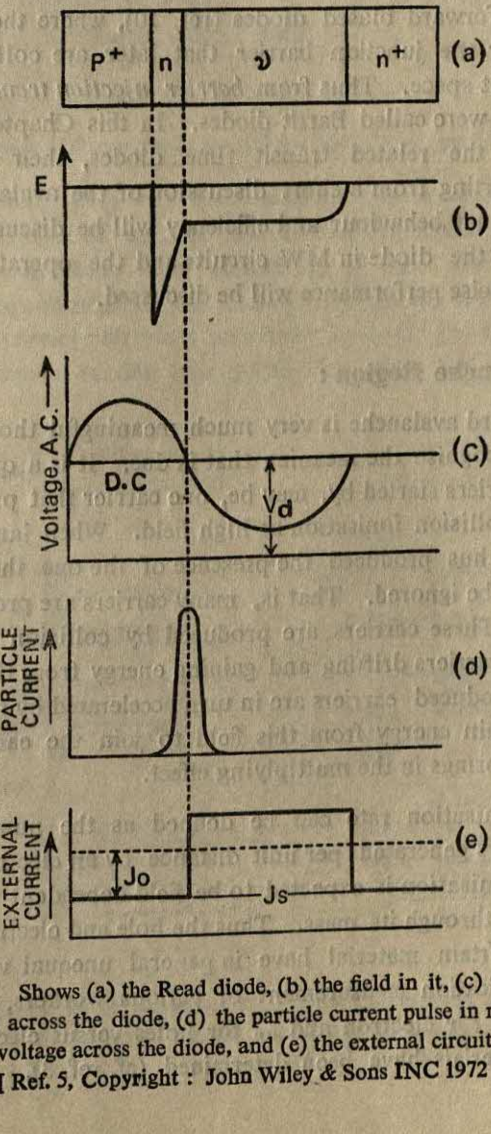


Fig. 12.0.1 Shows (a) the Read diode, (b) the field in it, (c) the voltage variation across the diode, (d) the particle current pulse in relation to the AC voltage across the diode, and (e) the external circuit current.
[Ref. 5, Copyright : John Wiley & Sons INC 1972]

form is expected to give the largest negative conductance. For this situation, it will be shown that the drift region transit angle $\omega T = 2\theta = \pi$ and the particle current lags the avalanching voltage by $\pi/2$ (cf. equation (12.3.15).

Later it was found by many workers (ref. 3, 5) that P-N junctions of any doping profile can exhibit this type of negative resistance. Further it was found that such negative resistance is also possible for forward biased diodes (ref. 10), where the carriers are injected over the junction barrier that later are collected after a certain transit space. Thus from *barrier injection transit time effect* these diodes were called Barrit diodes. In this Chapter, the Impatt diodes and the related transit time diodes, their principle of operation starting from a short discussion of the avalanche process, their large signal behaviour and efficiency will be discussed. Finally mounting of the diode in MW circuits and the operation principle therein and noise performance will be discussed.

12.1 Avalanche Region :

The word avalanche is very much meaningful though normally we fail to recognise the meaning that is due. It is a quick multiplication of carriers started by, may be, one carrier that produced more carriers by collision ionisation at high field. When large number of carriers are thus produced the presence of the one that started the process may be ignored. That is, many carriers are produced out of, as if none. These carriers are produced by collision ionisation of atoms with carriers drifting and gaining energy from the field. The ionisation produced carriers are in turn accelerated by the same field. Thus they gain energy from this field to join the earlier produced ones. This brings in the multiplying effect.

The ionisation rate can be defined as the number of hole electron pairs generated per unit distance by an electron or a hole. Hence the ionisation is expected to be field dependent as also on the particle type through its mass. Thus the hole and electron ionisation rates of a certain material have in general unequal values. From photomultiplication measurement of various workers, the ionisation coefficient α_n for electron and α_p for holes versus electric field for different materials have been determined (cf. ref. 3). It is seen that for limited field ranges around a field E_0 the ionisation coefficient α can be expressed as

$$\alpha = \alpha_0 (E/E_0)^m$$

where α_0 and m are constants, α_0 depending on the critical field E_0 while m varies between 3 to 9. A typical value for Si is 6. A better expression for the ionisation rate that fits over a wider field range is

$$\alpha = A \exp(-b/E)^n \quad \cdots \quad (12.1.1)$$

where in case of electrons in silicon at 300°K

$$A = 3{\cdot}8 \times 10^6 \text{ per cm.}, \; b = 1{\cdot}75 \times 10^6 \text{ V/cm, and } n = 1.$$

Continuing with our discussion on the avalanche process, a very large number of carriers are produced by avalanche multiplication. This must have been started by at least one carrier. The minimum requirement for the avalanche to occur, therefore is that, at least one carrier pair must have been produced per collision. Since ionisation causes carrier pair production, thus for a certain high

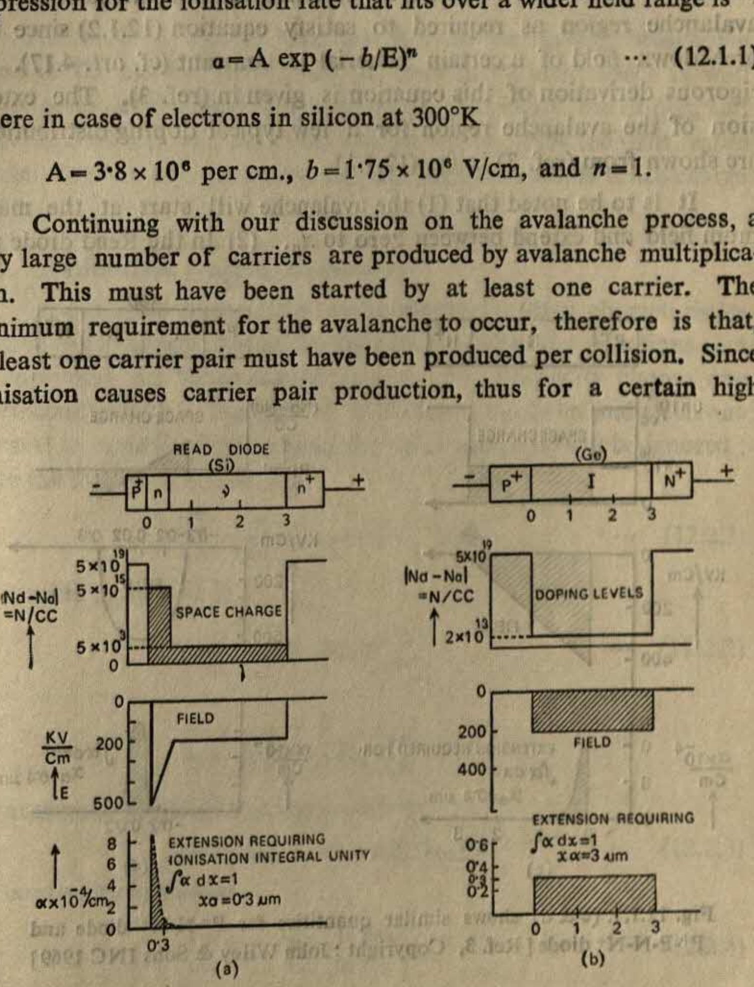


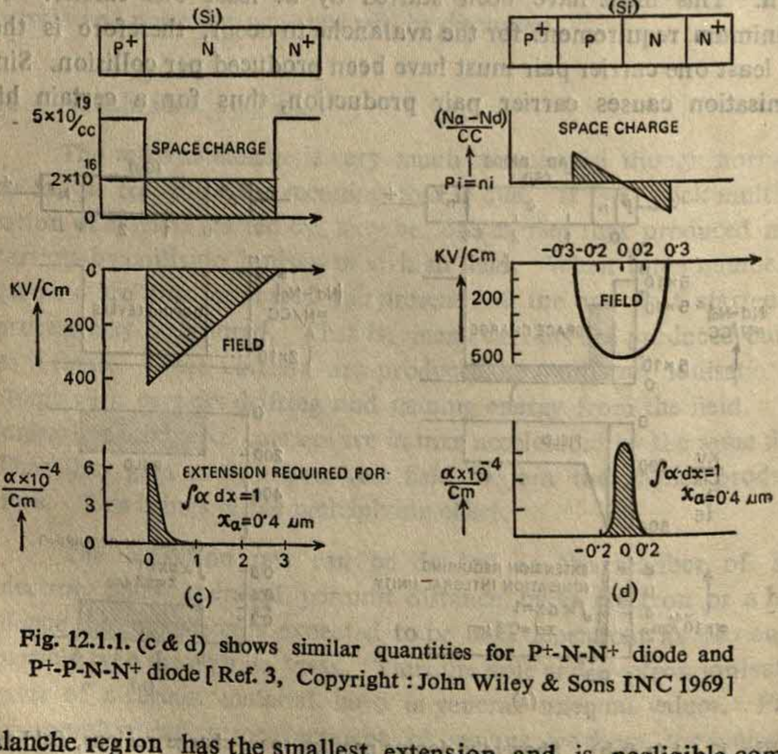
Fig. 12.1.1 (a & b). Shows the doping profiles, space-charge regions, field and the extent of the avalanche regions respectively for Read and P-I-N diodes. [Ref. 3, S. M. Sze. Copyright : John Wiley & Sons INC 1969]

field region we must have the minimum requirement for the avalanche to occur is

$$\int \alpha \, dx = 1 \qquad \cdots \quad (12.1.2)$$

This means that for a reverse biased junction of any type (either a P-N, P+-N-I-N+ or even for a met-sem. junction) when the doping profile is given there will have a certain minimum extension of the avalanche region as required to satisfy equation (12.1.2) since the breakdown field of a certain material is constant (cf. art. 4.17). A rigorous derivation of this equation is given in (ref. 3). The extension of the avalanche region for a few typical doping distributions are shown from (ref. 3) in fig. (12.1.1).

It is to be noted that (1) the avalanche will start at the maximum field and extend between zero to x_a. (2) In the Read diode the avalanche region has the smallest extension and is negligible compared to the total drifting space, the I-region, while in P-I-N diode the whole drift space is avalanche region. (3) The avalanche break-



(c) (d)

Fig. 12.1.1. (c & d) shows similar quantities for P+-N-N+ diode and P+-P-N-N+ diode [Ref. 3, Copyright : John Wiley & Sons INC 1969] down voltage depends on the detailed doping distribution, the material and the length of the drift region. Note also the length of the drift region in the diodes is of the order of 3 microns. If the carriers move through this space with saturated drift velocity, the transit time delay would be about 3×10^{-11} secs, double this time corresponds to 16 GHz. Later, we shall see that rigorous calculation gives a factor two less delay of the carriers.

The diode is biased such that it is in the reverse breakdown region of the characteristic when the field in the drift space is well above the saturation field of the material. Normally, the doping and dimensions are tailored such that with no applied bias the drift region mobile carriers are swept out and this region is totally space charge region.

12.2. Avalanche Region Equations

In order to analytically study the carrier flow characteristics through the avalanche region we need to formulate the equations that are applicable for the avalanche region. For simplicity choose one dimensional model. The applied field causes the holes and electrons to travel in opposite directions so the space charge cannot be ignored ; hence the Poisson's equation,

$$\frac{dE}{dx} = \frac{q}{\varepsilon_s \varepsilon_0} (N_d - N_a + p - n) \qquad \cdots \quad (12.2.1)$$

The current equations,

$$J_p = \mu_p \, q p E - D_p q \cdot \frac{dp}{dx} \qquad \cdots \quad (12.2.2)$$

and

$$J_n = \mu_n q n E + D_n q \cdot \frac{dn}{dx} \cdot$$

The avalanche generation at any location,

$$g = [\alpha_n(E)\,|J_n| + \alpha_p(p)\,|J_p|]/q \qquad \cdots \quad (12.2.3)$$

The continuity equations,

$$\frac{dp}{dt} = g - R - \frac{1}{q} \cdot \frac{\delta J_p}{\delta x} \qquad \cdots \quad (12.2.4)$$

and

$$\frac{dn}{dt} = g - R + \frac{1}{q} \cdot \frac{\delta J_n}{\delta x},$$

where R is the Shockley-Read thermal generation and recombination term (cf. Problem 24, Chapter 2).

12.3. The Read Diode :

In this section, we discuss the small signal theory of the Read diode which forms the basis for understanding more complex structures. Fig. (12.1.1a) shows the Read diode model. This diode can be regarded as consisting of three regions, the avalanche region extending from O to x_a followed by a drift region, the I-layer where the d.c. field remains above that needed for velocity saturation. The drift region ends in the high conductivity N+ substrate.

We assume the avalanche region boundaries with the *exit* saturation currents that are responsible for starting the avalanche as shown in fig. (12.3.1). It is assumed that from $x=0$ to the right the electron particle current increases while the hole particle current increases from $x=x_a$ to the left due to avalanche. The reverse saturation genaration currents J_{ps} and J_{ns} shown at the boundaries are some total generation of the whole region $x=0$ to $x=x_a$. More accurately, these are little less because the avalanche zone is smaller in size than the total depletion region.

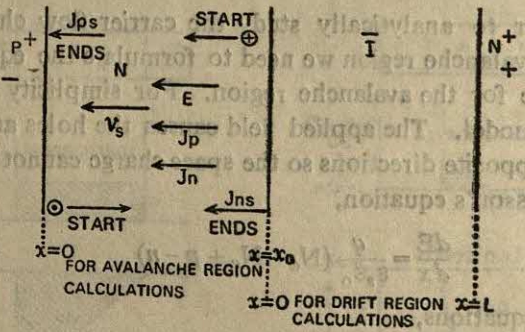

Jps
START
P+
ENDS
N
E
Vs
Jp
Jn
Jns
START
ENDS
I
N+
x=0
FOR AVALANCHE REGION CALCULATIONS
x=xa
x=0 FOR DRIFT REGION CALCULATIONS
x=L

Fig. 12.3.1. Shows the different currents at start and end of multiplication, also the two regions separately.

To make the avalanche region properties tractable analytically a few simplifications to the equations of article (12.2) can be made and these are :

(1) carrier diffusion is ignored. Though this has no justification except that the carrier velocity at such high fields are always

saturated but that does not rule out diffusion entirely, but this, simplifies the analysis. (3) Equal ionisation rates $a = a_n = a_p$ for holes and electrons are assumed. (3) Both hole and electron carrier velocities are assumed saturated and equal to v_s. (4) Lastly, the thermal generation recombination term R is ignored. With these simplifications, we have instead the following equations as can be understood with reference to fig. (12.3.1) for the sign conventions.

$$J_p = qp\, v_s \,;\; J_n = qn\, v_s \qquad \cdots \qquad (12.3.1)$$

$$\text{and} \quad J = J_n + J_p$$

From equations (12.2.4) and equation (12.3.1) with $R = 0$ and $a_n = a_p = a$

$$\frac{1}{v_s}\left(dJ_p/dt\right) = -\left(dJ_p/dx\right) + aJ \qquad \cdots \qquad (12.3.2)$$

$$\frac{1}{v_s}\left(\frac{dJ_n}{dt}\right) = (dJ_n/dx) + aJ$$

The Poisson's equation becomes

$$\varepsilon\varepsilon_0 \frac{dE}{dx} = q(N_d - N_a) + (J_p - J_n)/v_s \qquad \cdots \qquad (12.3.3)$$

When we have from equation (12.3.2)

$$\frac{1}{v_s}\frac{d}{dt}(J_p - J_n) = -\frac{d}{dx}(J_p + J_n) \qquad \cdots \qquad (12.3.4)$$

which with equation (12.3.3) gives

$$\frac{d}{dx}\left(\varepsilon\varepsilon_0 \frac{dE}{dt} + J_p + J_n\right) = 0,$$

as is expected since for an area A the total current

$$I = A\left(\varepsilon\varepsilon_0 \frac{dE}{dt} + J_n + J_p\right) \text{ should not have a gradient.}$$

Adding equations in (12.3.2) and integrating for the extension of the avalanche regions with the avalanche region transit time

$$T_a = (x_a/v_s), \text{ we have}$$

$$T_a \cdot \frac{dJ}{dt} = -\left\{J_p - J_n\right\}_0^{x_a} + 2J \int_0^{x_a} a\, dx \qquad \cdots \qquad (12.3.5)$$

Remembering that the potential on reverse bias is continuously increasing from left to right at the boundary at $x = 0$ the current $(J_p - J_n) = (2J_{ps} - J)$ where J_{ps} is the reverse bias thermal generation hole current going out on the left. Similarly, at $x = x_a$; $(J_p - J_n) = J - 2J_{ns}$ where J_{ns} is the thermal generation electron current entering (particles going out) at $x = x_a$. With these boundary values equation (12.3.8) reduces to

$$\frac{dJ}{dt} = \frac{2J}{T_a}\left(\int_0^{x_a} \alpha\, dx - 1\right) + \frac{2J_s}{T_a} \qquad \cdots \qquad (12.3.6)$$

where $J_s = J_{ps} + J_{ns}$, which is small compared to the avalanche produced total conduction current J. Hence equation (12.3.6) can be written as

$$\frac{dJ}{dt} = \frac{2J}{T_a}\left(\int_0^{x_a} \alpha\, dx - 1\right) \qquad \cdots \qquad (12.3.7)$$

Moreover, the particle current that is going out of the right-hand boundary of the avalanche region is electron current since the holes travelling to the left, reaching the P-doped region where they are majority carriers hence have comparatively very short transit times. Hence $J = J_0 + J_n(t)$ where J_0 is the avalanche produced d.c. bias current and $J_n(t)$ is the time dependent part of a.c. electron particle current that enters the drift region (cf. 12.3.1) at x_a.

Ignoring space charge hence field gradient in the avalanche region, we write α around the d.c. value α_0 as $\alpha = \alpha_0 + \left(\frac{d\alpha}{dE}\right)_{\alpha_0} \cdot E_1$ where E_1 is the incremental avalanche region field and is equal to (v_1/x_a) with v_1 as the incremental voltage across the avalanche region. To a first order of approximation from equation (12.3.7) the electron particle current entering the drift region can be written as*

$$\frac{dJ_n(t)}{dt} = \frac{2J_0\, v_1 \overline{\alpha_0'}}{T_a} \qquad \cdots \qquad \cdots \qquad (12.3.8)$$

where $\overline{\alpha_0'} = \left(\int_0^{x_a} \frac{d\alpha}{dE}\right)_{\alpha_0} dx/x_a$, since under avalanche condition

---

* It is to be noted that the d.c. arising out of the nonlinearities that are ignored in framing equation 12.3.8 above must surely cancel out. Taking the next higher order terms $(2\, J_n\, v_1\, \overline{\alpha_0'})/T\alpha$ and $(2\, J_0\, \overline{\alpha_0''}\, v_1{}^2)/x\alpha$. $T\alpha$ where $\overline{\alpha_0''} = \frac{1}{x\alpha} \int_0^{x_a} (d^2\alpha/d^2E)_{\alpha_0}\, dx$, the rectified d.c. from these two terms cancels out because the second derivative of α is negative as discussed under large signal effects (cf. art. 12.6) while $\overline{\alpha_0'}$ is signal amplitude v_1 dependent.

$\int_0^{x_a} \alpha_0 \, dx = 1$. It is to be noted that the avalanche region width x_a decreases with increased field hence increased current but for small signal we speak of an average width x_a.

For the drift region on the other hand we assume that at the positive peaks of the a.c. cycle across the diode the drift region field remains low enough not to cause any ionisation in this region while at the negative peak it remains high enough not to let the carrier velocity drop below saturation velocity v_s. Secondly, in calculating the terminal current, we need to remember that the electron particle current that enters the drift region is a current distributed in space and time. Hence to find the total current we consider

$$\mathrm{I} = -\mathrm{A}\left[\mathrm{J}_n(x, t) + \frac{\varepsilon\varepsilon_0 d\mathrm{E}}{dt}\right] \quad \cdots \quad (12.3.9)$$

Integrating this over the drift region that extends from, say, $x=0$ to $x=\mathrm{L}$ we have,

$$\mathrm{I} = \mathrm{A}\left(\mathrm{J}_c + \frac{\varepsilon\varepsilon_0}{\mathrm{L}} \frac{d\mathrm{V}}{dt}\right) \quad \cdots \quad (12.3.10)$$

$$\text{where } \int_0^{\mathrm{L}} -\mathrm{E}\, dx = \mathrm{V} \text{ and } \mathrm{J}_c = -\left[\int_0^{\mathrm{L}} \mathrm{J}_n(x, t)\, dx\right]/\mathrm{L} \quad \cdots \quad (12.3.11)$$

Note $x=0$ to $x=\mathrm{L}$ for drift region's calculations in fig. (12.3.1). That is, the block of electron charge gives a conduction current J_c to which the capacitive current needs to be added to give the total current which remains of constant value so long the particles are on transit through the drift space.*

In the drift region the particle velocity is constant. Hence the particle current for a certain charge element at time t and position

---

*This can also be viewed in terms of induced charges. As a lump of charged particles approaches the electrode more field lines get terminated on it. Taking the simple case of plane parallel electrodes by Gauss' theorem the charge element δq must be accounted for by a field difference as $(\mathrm{E}_2 - \mathrm{E}_1)$ where E_1 and E_2 are the fields on either side on the charge element. While for the voltage drop across the electrode $\mathrm{V} = \mathrm{E}_1 x + \mathrm{E}_2(\mathrm{L} - x)$. This can be written as $\delta\mathrm{V} = \mathrm{L}\delta\mathrm{E}_2 - \delta q \delta x/\varepsilon\varepsilon_0$ and hence $(\varepsilon\varepsilon_0\, \delta\mathrm{V}/\delta t)/\mathrm{L} + (v\, \delta q) = (\delta\mathrm{E}_2/\delta t)$. Thus we see that the induced current on the electrode $(\delta\mathrm{E}_2/\delta t)$ is the sum of the capacitive current and the conduction particle current.

x is the same as it was at $x=0$ at a time (x/v_s) earlier. Diffusion being neglected as mentioned earlier. That is

$$J_n(x, t) = J_n(0, \overline{t - x/v_s}) = J_n\left(t - \frac{x}{v_s}\right) = J_n(t_1) \qquad \cdots \quad (12.3.12)$$

hence equation (12.3.11) can be written with $T = (L/v_s)$ as

$$J_c = \left(\int_{t-T}^{t} J_n(t_1)\, dt_1 \right) / T \qquad \cdots \quad (12.3.13)$$

Assuming that the electron particle current entering the drift region at $x=0$ is of the form $J_1 \exp(j\omega t)$, we have from equation (12.3.13) the conduction current

$$J_c = J_1 \frac{\sin\theta}{\theta} \exp j(\omega t - \theta) \qquad \cdots \quad (12.3.14)$$

where θ is half of the transit angle ωT. This shows that the conduction current phase delay instead of being ωT is reduced and as if it due to half of the drift time. Thus we can consider the total drifting charge to be located at the centre of the distributed charge.

Using $J_n(t) = J_1 \exp(j\omega t)$ and $v_1 = V_a \exp(j\omega t)$ in equation (12.3.8) we have

$$J_1 = [b V_a \exp(-j.\pi/2)] \qquad \cdots \quad (12.3.15)$$

$$\text{where} \quad b = (2J_0 \overline{a_0'})/\omega\, T_a \qquad \cdots \quad (12.3.16)$$

Thus we see that the particle current lags the avalanche region voltage by $\pi/2$.

Equation (12.3.14) with equation (12.3.15) gives

$$J_c = \frac{J_0 \overline{a_0'}}{(\omega T_a/2)} V_a \frac{\sin\theta}{\theta} \exp - j(\theta + \pi/2) \exp(j\omega t) \qquad \cdots \quad (12.3.17)$$

That is the conduction current delay over the avalanche region voltage is due to the sum of the particle current phase delay of $\pi/2$ and transit time delay of $\theta = \omega T/2$.

Now we are poised for an overview of the whole situation. Starting with the avalanche region, the total current out of the avalanche region, with C_a as the capacitance of this region is obtained on using $J_1 = -jbV_a$ from equation (12.3.15), as

$$J = -jb\, V_a + j\omega\, C_a\, V_a/A \qquad \cdots \quad (12.3.18)$$

$$\text{Hence} \quad (J_1/J) = 1/(1 - \omega^2/\omega_r^2) \qquad \cdots \quad (12.3.19)$$

where $$\omega_r^2 = (\omega b A/C_a) = \frac{2J_0 A\overline{\alpha_0}'}{T_a C_a} = \frac{1}{L_r C_a} \quad \cdots \quad (12.3.20)$$

That is the avalanche region acts an inductor $L_r = (T_a/2J_0 A\overline{\alpha_0}')$ which with the capacitor C_a of this region forms an antiresonant circuit. The conduction current flowing through the avalanche region is the inductor current while the total current is the external current of this antiresonant circuit. The antiresonant frequency being given by equation (12.3.20).

For the drift region on the other hand the total current can be written with drift region capacitance C_d and using equation (12.3.14) as

$$J = J_0 + j\omega C_d V_d/A = (J_1 \frac{\sin\theta}{\theta} \exp - j\theta) + j\omega C_d V_d/A \quad \cdots \quad (12.3.21)$$

which gives the voltage V_d across the drift region in terms of (J_1/J) which, in turn, is given by equation (12.3.19). Similarly, the avalanche region voltage can be obtained from equation (12.3.15) with equation (12.3.19)). Thus, the diode admittance can be obtained which on simplification gives for $Y = AJ/(V_a + V_d)$ as :

$$Y = G + jB = \frac{j\omega C_d (1 - \omega^2/\omega_r^2)}{[1 - (\omega^2/\omega_r^2)(1 + C_d/C_a) - \sin\theta\,(\exp - j\theta)/\theta]} \quad \cdots \quad (12.3.22)$$

Noting that in separating the real and imaginary parts, the common denominator will be a sum of the square of two terms. We take the numerators of the real and imaginary parts of Y from this equation when we have for

$$\mathrm{Nu}(G) = -2[(\omega^2/\omega_r^2) - 1]\, C_d \sin^2\theta/T \quad \cdots \quad (12.3.23)$$

and

$$\mathrm{Nu}(B) = j\omega\, C_d(1 - \omega^2/\omega_r^2)\,[1 - (\omega^2/\omega_r^2)(1 + C_d/C_a) - \sin 2\theta/2\theta] \quad \cdots \quad (12.3.24)$$

From equation (12.3.23) we note that (i) for all values of transit time delay negative conductance is presented by the diode only above the resonant frequency ω_r, (ii) for the same diode that is constant transit time or $2\theta_r = \omega_r T$, the turning points of the numerator of the conductance occur at

$$\tan\theta = (\theta_r^2 - \theta^2)/\theta \quad \cdots \quad (12.3.25)$$

Negative resistance occurs only for ω greater than ω_r that is θ being positive greater than θ_r ; hence for negative value of tan θ. The first branch of negative value of tan θ starts at $\pi/2$ where it decreases very rapidly with θ hence a maximum of negative conductance is obtained just ahead of the twice transit angle of $2\theta = \omega T = \pi$ as shown in fig. (12.3.2). Further ahead of this transit angle at $\omega T = 2\pi$ that is at $\theta = \pi$, equation (12.3.23) shows that the conductance becomes zero. $2\theta - \pi$ to 2π is about the first negative conductance branch that is useful for practical purposes. It can be easily seen from equation (12.3.23) that for ω very much greater than ω_r the magnitude of the negative conductance decreases being inversely proportional to w^6 and that for infinitely large transit angle that is large ω equation (12.3.22) shows that the admittance tends to be capacitive only. It is to be remembered that in the above discussion the bulk substrate resistance has been ignored which is to be considered in series with the parallel combination of diode conductance and susceptance.

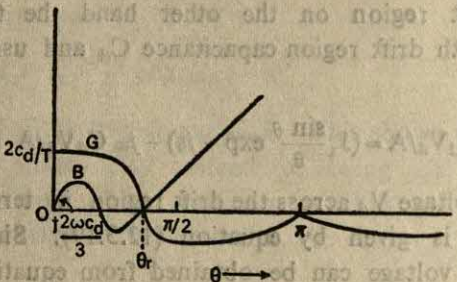


Fig. 12.3 2. Shows the variations of conductance (G) and susceptance (B) versus the transit angle θ.

Before we discuss about the detailed nature of the susceptance variation it is worth considering the phasor diagrams of the different current voltages before and after the antiresonant frequency ω_r. Starting with the avalanche region voltage V_a as reference from equation (12.3.15) we see that J_1 (the particle current entering the drift space) lags V_a by $\pi/2$. This is depicted in figs. (12.3.3a) and (12.3.3b) for below and above ω_r respectively. From equation (12.3.14) we have J_0 the drift region conduction current lags J_1 by θ as also is reduced by the factor (sin θ/θ). From equation (12.3.19) the total current is either in-phase or of opposite phase before and after the antiresonant frequence ω_r respectively to J_1. Not only this, since we are considering around antiresonance, J is smaller than J_1. Using equation (12.3.21) we can find the drift region capacitive current $(j\omega C_d V_d/A)$. Note that in both cases this

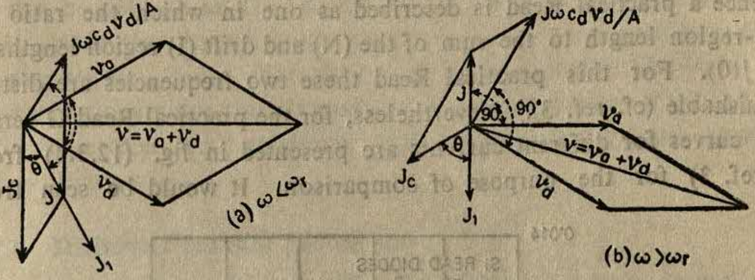


Fig. 12.3.3. Shows the phasor relations between the different currents and voltages in a Read diode above and below ω_r.

is in the first quadrant. V_d is then drawn lagging this current by $\pi/2$ and so the total applied voltage $V = V_a + V_d$ is drawn as shown. Note that in both cases V is in the fourth quadrant. While for $\omega > \omega_r$ only the angle between v and J is greater than $\pi/2$. That is negative resistance is obtained only for the frequencies greater than the antiresonant frequency. The series bulk substrate resistance would give a voltage drop in phase with J. In either case this will bring the total voltage including the resistance drop more towards J. For frequencies above antiresonance, the negative conductance will decrease since the total voltage tends more towards perpendicular to J.

The variation of the susceptance B on the other hand can be understood with the help of equation (12.3.24). For ω very much less than ω_r from equation (12.3.22) it can be seen to be capacitive and decreasing with increase of ω. For ω a little lower than ω_r the susceptance is inductive as seen from equation (12.3.24) so long

$$\left[\frac{\theta^2}{\theta_r^2}\left(1+\frac{C_d}{C_a}\right)+\frac{\sin 2\theta}{2\theta}\right]>1$$

and remains inductive till at $\omega = \omega_r$ it is again zero and remains capacitive thereafter as shown in fig. (12.3.2). From the above discussions it is apparant that (i) for most of the useful range of negative conductance the susceptance is capacitive (ii) that for a certain current hence a certain ω_r (c.f. equation 12.3.20) one zero of susceptance occurs at a frequency a little lower than ω_r and the other at ω_r. In the literature (cf. ref. 3) these two frequencies in theoretical Read has not been separately described. Since experimentally it is difficult to construct a negligibly narrow N-region hence a practical Read is described as one in which the ratio of N-region length to the sum of the (N) and drift (I) region lengths is (1/10). For this practical Read these two frequencies are distinguishable (cf. ref. 3). Nevertheless, for the practical Read G versus B curves for different currents are presented in fig. (12.3.4) from (ref. 3) for the purpose of comparison. It would be seen from

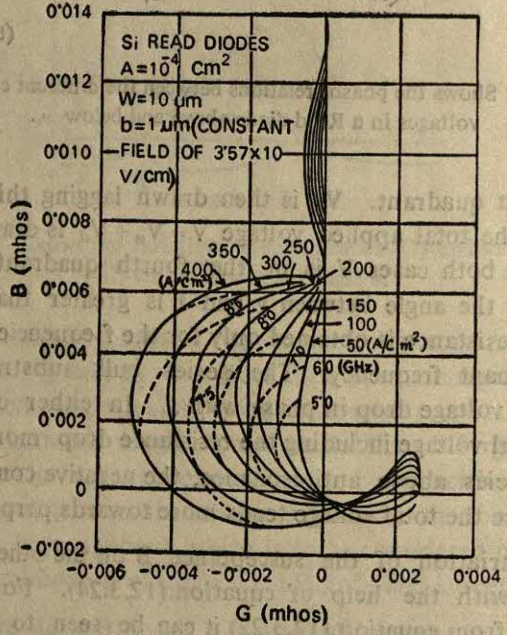


Fig. 12.3.4. Shows the B versus G relation for a practical Read diode [Ref. 3, Gummel and Scharfetter John Wiley and Sons INC 1969]

this figure that in general the zero conductance and zero susceptance frequencies are different. Hence, we define two frequencies, namely, the resonant frequency f_r is the frequency at which the susceptance B changes from inductive to capacitive and the cut-off frequency f_c is the minimum frequency at which the conductance G changes from positive to negative value. Obviously, it is clear that both of these frequencies refer to the avalanche region anti-resonant frequency ω_r that we have come across in the theoretical Read. Hence its proportionality to $\sqrt{I_0}$ should be bourne in mind —cf. equation (12.3.20). It is to be noted that the diode negative Q can be defined as $Q=(B/G)=-\omega CR$ since it is a parallel circuit. Also that its minimum occurs at about the transit time frequency where $\omega T=\pi$. This frequency is the optimum for build up of oscillations since this depends on the factor $\exp(-\omega t/Q)$ and Q here is negative and minimum that is required for faster growth of oscillations.

12.4 Different Impatt Structures

The analysis of the different Impatt structures shown in fig. (12.1.1) are more complex than the ideal Read structure discussed in the earlier article. In them the avalanche region is not short enough to enable one to ignore the additional delay due to the avalanche region transist time. That is neither the avalanche region space charge nor the transit delay of carriers through this space charge can be ignored. For such diodes, therefore, the pertinent equations (cf. Art. 12.2) lead to a set of nonlinear equations. The solutions of these are obtained by computer calculations (cf. ref. 3). From the results of such calculations a few general conclusions can be made :—

(1) Under proper biasing conditions junction diodes of various doping profiles can give rise to incremental negative conductance.

(2) The general nature of the admittance plot as shown for an N^+-N-P^+ diode in fig. (12.4.1) is not very much different from that of the Read diode, shown in fig. (12.3.4) except for the absolute values.

(3) Both the cut-off frequency f_c and the resonant frequency f_r are in general different but both are proportional to the square root of the direct current as in the Read diode.

(4) At a certain current density both of these frequencies decrease with the increase of the avalanche region width.

(5) For a given total depletion region width as the avalanche region width increases, the negative conductance frequency band increases but the negative conductance value decreases. In the limi-

ting case, when the entire drift space is the avalanche region (cf. fig. 12.1.1 P-I-N diode), the cut-off frequency becomes zero. That is the negative conductance is obtained from very low frequency upwards.

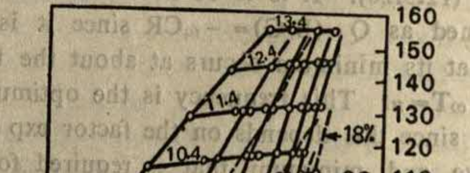


Fig. 12.4.1. Shows the B versus G curves for an N^+-N-P^+ diode [Ref. 4, Scharfetter and Gummel, Copyright © 1969 IEEE TRANS ED 16, No. 1]

(6) A narrow depletion width obviously then gives a lower Q and since the transit time through the diode decides the operating frequency a narrow total depletion width is required for high frequency diodes.

12.5 Read diode efficiency

For efficient operation of a Read diode, as the carriers move through the drift region, it is required to deliver as large a current pulse as possible to the external circuit centred around the voltage minimum across the diode (cf. fig. 12.1). Moreover, this is required to be achieved without a reduction of electric field below that required to maintain velocity saturation when the charge is on transit through the drift region. A large current pulse requires a large a.c. voltage to give the required large ionisation produced avalanche current. A large a.c. voltage amplitude across the diode, on the other hand, reduces the diode voltage during the later part of the transit of the charge in the negative cycle. For a square wave external circuit current pulse the external circuit develops the fundamental sinusoidal voltage. It may be surmised, therefore, that as a compromise, the maximum a.c. voltage amplitude that can possibly be developed across the external circuit is about half of the average drift region voltage drop V_d as shown in fig. (12.1). While the maximum external circuit square shape current pulse swing can be from J_s the reverse saturation current to $(J_s + 2J_0)$ where J_0 is the avalanche produced time independent d.c. current (cf. fig. 12.1). Under these conditions the d.c. power input can be written as

$$P_{in} = (V_a + V_d)(J_s + J_0) \quad \cdots \quad (12.5.1)$$

where V_a is the average avalanche region voltage and J_s is the reverse bias saturation current.

While the a.c. power output is given by

$$P_{out} = (4J_0/\pi\sqrt{2})(V_d/2\sqrt{2}) \quad \cdots \quad (12.5.2)$$

Hence the efficiency

$$= (P_{out}/P_{in}) = (1/\pi).\ (V_d/\overline{V_a + V_d})\ (J_0/\overline{J_s + J_0}) \quad \cdots \quad (12.5.3)$$

Note that the power contribution from the avalanche region has been ignored. This is because at the operating frequency the avalanche region behaves as capacitance to the external current (cf. fig. 12. 3. 3b) and hence do not contribute to the fundamental R.M.S. power.

Equation (12.5.3) shows that Read diode efficiency would be largest if V_d and J_0 are very much large compared to V_a and J_s respectively when the ideal efficiency is 30%. For a practical Read diode computer results give for semiconductors like Ge and GaAs with nearly equal ionisation rates of holes and electrons $(V_a/V_d) \simeq 1/3$ hence an ideal efficiency of 23%. While for silicon holes and electrons have unequal ionisation rates that give $(V_a/V_d) = 1$ and has an ideal efficiency of 15%.

12.6 Large signal Considerations

A number of authors have given large signal analysis of Impatt oscillators but the most accurate and complete one is again a computer solution. Fig. (12.6.1) shows (ref. 4) the time evolution of the electric field and hole and electron densities at 1/4 cycle intervals as a function of position in a Read (P⁺N-I-N) diode analysed by Scharfetter and Gummel. With it a phase plot of terminal current voltage is also shown. From fig. (12.6.1) it is to be noted that (*a*) the electrons and hole generation begins when the terminal voltage is around the maximum (*b*) a quarter of a cycle later the charge pulses are fully formed and drift into the respective directions, the holes disappear quickly from the active region to the left while electrons to the right entering the drift space. (*c*) The electrons drift through the drift space for about half a cycle as particle current during the negative excursion of the terminal voltage. (*d*) In the next quarter of a cycle the terminal R.F. voltage is positive and increasing when the last bit of electron charge is delivered to the anode. The

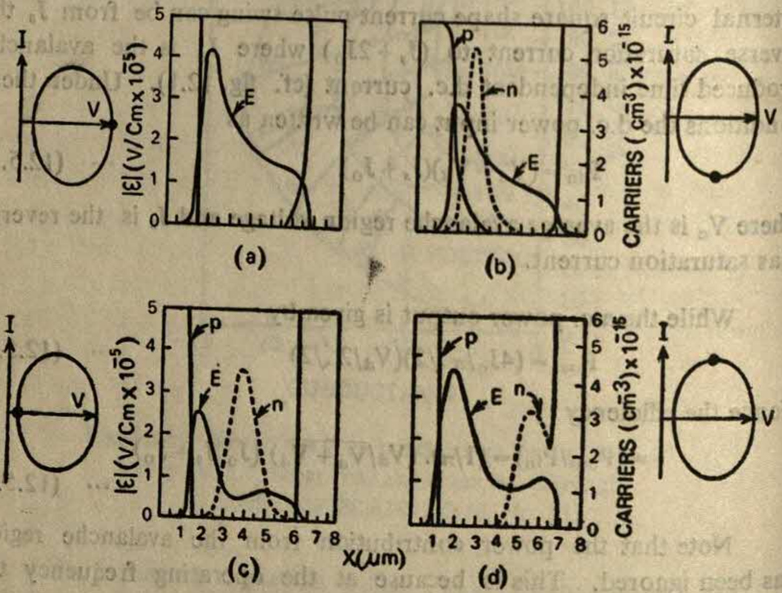


Fig. 12.6.1. Shows the time evolution of the electric field, hole and electron densities at ¼ cycle intervals Ref. 4 [by Gummel and Scharfetter by permission, Copyright © IEEE TRANS ED 16, 1969]

diode remains reverse biased always while the displacement current is large and positive occurring during the first and last quarter cycles. The computer calculations indicate that under large signal conditions the variation of electric field and space charges are of large magnitude bringing with it nonlinearities that give rise to self-limiting of the avalanche process. Otherwise, increased avalanche in the positive half cycle would have given increased residual current at each cycle that would have lead to unlimited d.c. current increase as the oscillations grow to large amplitudes. A few other effects, such as, lowering of the field disproportionately in the diode body at different times during the transit of the charge pulse is to be noted from the above figure. These non-linearities are due to the space charges and have profound effect on the carrier transit time ; hence, efficiency through its dispersal due to diffusion leading to different external current pules shape. For example, (*i*) the space charge field reduces the total field and may turn off the avalanche prematurely and thus can reduce the phase delay provided by the avalanche. This, in turn, reduces the negative conductance of the diode. (ii) When the drifting carriers are in transit the space charge depresses the field to the left of the carrier pulse. Thus the carrier drift velocity may drop below the saturation velocity when some of the carriers (cf. fig. 12.6.1c) may be very slowly moving since they are in the low field region (trapped). This does sufficiently change the terminal current waveform so that the power generated at the transit time frequency range is reduced.

In the estimation of efficiency we have seen that an increase of the reverse saturation current I_s would decrease the efficiency. The reverse saturation current can increase due to (i) the avalanche region is very narrow and the generated carrier concentration is high. Hence, there will have a large carrier density gradient giving rise to large diffusion current in both directions of the carrier packet. Of the carriers that have gone due to diffusion in the opposite direction some may come back later in the cycle due to the field to add to the tail of the main current. This current would, therefore, behave as a lossy reverse saturation current. This may even go as far as to reduce the phase delay in the next oscillation cycle thus decreasing the efficiency. (ii) Some experiments have shown that poor injection efficiency of a contact can inject minority carriers in heavily doped region causing an increase of saturation current leading to decrease of efficiency.

The non-linearity in field versus ionisation rate may directly cause efficiency reduction. Such as in very narrow diodes the field required for avalanche breakdown may be very high. While for very high field ($d\alpha/dE$) decreases. Thus less current modulation is obtained for a given R.F. voltage and the efficiency is reduced.

The above discussion shows that increased R.F. voltage in general is expected not to give proportional increase of R.F. terminal current maximum. Obviously, then negative conductance will decrease with increased R.F. swing. This is shown in fig. (12.4.1) from computational results of Gummel and Sharfetter. Note that in general the diode susceptance increases with increase of R.F. voltage swing.

On the other hand, under large signal conditions at the point of maximum efficiency the diode negative conductance becomes very small. Hence any extrinsic series resistance that was so far ignored in the previous discussions will have appreciable effect on efficiency. Such excess series resistance may arise due to causes like (i) unswept epitaxial or I-layer (cf. ref. 6), that may remain due to faulty biasing or design (ii) the skin effect causes non-uniform current distribution in the substrate thus effective substrate resistance is increased causing a decrease of efficiency with increase of frequency.

Moreover, there is inherent large signal power output limitation as related to the frequency and diode reactance due to the following reasons. Large power output implies large power input from which it has to be derived. But maximum power input is limited by the breakdown voltage hence the breakdown field and the diode length, while the maximum current that can be passed through the diode depends on the saturation velocity and the diode cross-sectional area. The maximum diode breakdown voltage is obtained when the avalanche ionisation caused breakdown field E_c in obtained throughout the diode length.

That is $V_m = E_c.L$ ... (12.6.1)

while $I_m = qn\,v_s A$ ... (12.6.2)

But since the current would be space charge limited the maximum number of charge carries that can takes part in the process is limited by the Poisson's equation

$$\frac{dE}{dx} \simeq (E_c/L) = (qn/\varepsilon\varepsilon_0) \quad \cdots \quad (12.6.3)$$

Combining equations (12.6.1) to (12.6.3) we have for the maximum power that can possibly be imparted to the mobile carriers is

$$P = V_m I_m = \varepsilon\varepsilon_0 E_c^2 \, v_s \, A \quad \cdots \quad (12.6.4)$$

Remembering that the diode reactance is given by

$$X = (1/2\pi fC) = (L/2\pi f\varepsilon\varepsilon_0 A) \quad \cdots \quad (12.6.5)$$

and that the transit time frequency f is given by $\omega T = \pi$ that is

$$f = (v_s/2L) \quad \cdots \quad (12.6.6)$$

we have from equation (12.6.4) with (12.6.5) and (12.6.6)

$$P.X = (E_c^2 \, v_s^2 / 4\pi f^2) \quad \cdots \quad (12.6.7)$$

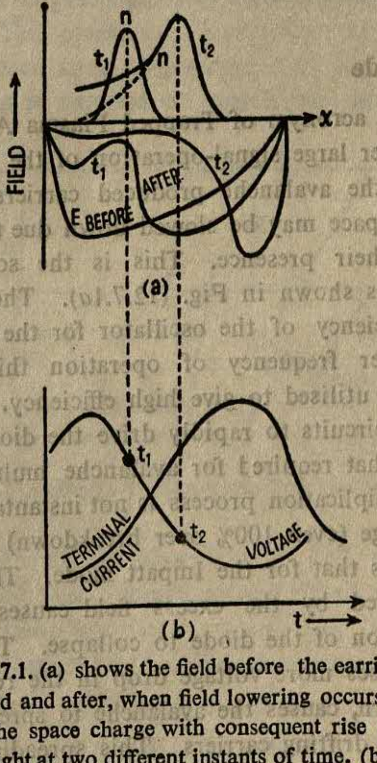


Fig. 12.7.1. (a) shows the field before the earriers are produced and after, when field lowering occurs on the left of the space charge with consequent rise of field on the right at two different instants of time. (b) shows the diode voltage and current at these instants.

For large signals the susceptance is large compared to the conductance; hence the total admittance or the impedance is limited by the reactance. In all high frequency high power semiconductor devices we would like to keep the circuit losses to low value hence the impedance has to be limited to a minimum value. In such cases we see from equation (12.6.7) that the maximum power that can be imparted to the carriers decreases as reciprocal of the square of the operating frequency and that the $P.X.f^2$ product is limited by the semiconductor parameters. At present in the 40—150 GHz range the Impatt diodes are usuable for waveguide communication systems and as pump sources for parametric amplifiers. Diodes, with both type of carriers used to drift through respective drift spaces on either side of the avalanche region, called double drift Impatts have been realised. With these CW powers as high as 380 *m*W at 92 GHz at 12·5% efficiency has been realised while at 150 MHz 150 *m*W output power has been obtained.

12.7 Trapatt mode

This is the acronym of Trapped Plasma Avalanche Transit Time mode. Under large signal-operation of the Impatt oscillator it was noted that the avalanche produced carriers while drifting through the drift space may be slowed down due to the field reduction caused by their presence. This is the so-called Trapped Plasma effect and is shown in Fig. (12.7.1*a*). Though this effect decreases the efficiency of the oscillator for the transit time frequencies, for lower frequency of operation this effect of field suppression can be utilised to give high efficiency. It is possible in high Q Impatt circuits to rapidly drive the diode to a voltage well in excess of that required for avalanche multiplication. This is because the multiplication process is not instantaneous. Because of the excess voltage (even 100% over breakdown) the multiplication greatly exceeds that for the Impatt mode. The large amount of carriers produced by the excess field causes the field in the maximum field region of the diode to collapse. This local increased conductivity causes more voltage drop in the rest of the drift region which, in turn, causes the avalanche to spread little more in the direction of the drifting carriers. This spreading of avalanche with field suppression on the left of it and field increase on the right continues, ultimately leaving the diode filled with plasma and with a very small voltage drop across the diode. If the terminal current is maintained the plasma is removed from the diode after a certain recovery time at the end of which the diode voltage returns to its breakdown bias value and the current returns to its low bias value. The field profile carrier densities and the terminal current, voltage are shown in fig. (12.7.1) at a specific time as marked in the figure. The time required to remove the plasma depends on the density of the plasma and the magnitude of the recovery current which, in turn, is circuit controllable. However, a recovery time of 5—10 times the transit time of carriers moving at the saturated drift velocity is typical.

A simple circuit for realising the Trapatt mode of operation consists of a transmission line with the diode mounted at one end and a low pass filter at the other end as shown in fig. (12.7.2). The low pass filter is placed approximately quarter wavelength away from the diode. This presents a short circuit to all the harmonics of the Trapatt oscillation frequency. On the load side of the L.P.F. some tuning is required to match the diode impedance to the load impedance. The inductance (L_s) and shunt capacitance (C_s) are the series lead inductance and shunt capacitance of the diode. The capacitance may also include any additional capacitance due to Impatt tuners or distributed capacitance near the diode.

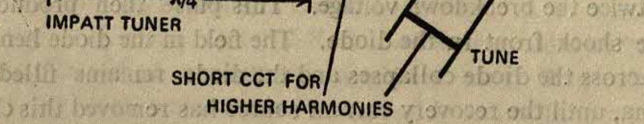


Fig. 12.7.2. Shows the Schematic Trapatt Circuit.

Fig. 12.7.3 shows the diode conduction current and voltage waveforms. Note that the diode voltage rapidly drops to a value near zero volts. This negative going voltage pulse propagates down the line to the short circuit provided by the low pass filter. The short circuit gives minus one reflection coefficient and causes a voltage pulse to be reflected back towards the diode. This positive voltage pulse adds to the diode bias voltage producing a momentary applied pulse of twice the breakdown voltage. This pulse then produces an avalanche shock front in the diode. The field in the diode hence the voltage across the diode collapses and the diode remains filled with the plasma, until the recovery current source has removed this charge. The diode again tries to oscillate in the Impatt mode. With proper tuning the Impatt frequency grows to initiate the process. When the Impatt oscillation amplitude has grown enough it charges the depletion layer capacitance to a voltage about twice the breakdown voltage. This, in turn, initiates the travelling avalanche zone which drops the voltage to almost zero. The particle current is very high during the low voltage state. After the plasma has been removed from the diode the current drops to low average value and the voltage returns to the breakdown value. The operation of the diode is as if Impatt frequency is pulse operating at the Trapatt

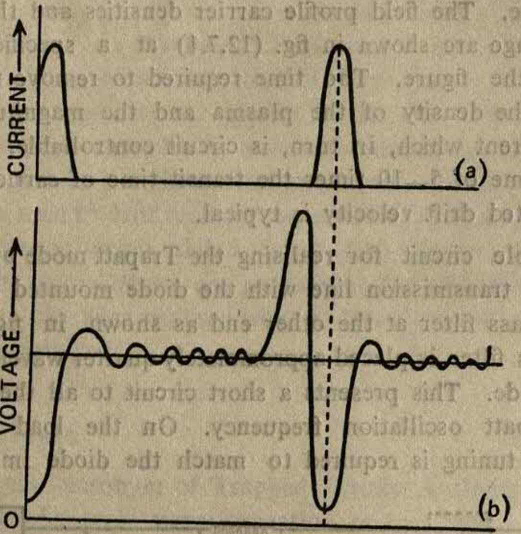


Fig. 12.7.3. Shows (a) the diode conduction current (b) the diode in Trapatt mode.

frequency. Since the voltage is low when the current is high and vice versa the computed efficiency is high about 60% for silicon diodes and 50% for Ge diodes. A typical MW circuit is shown in fig.(12. 7.4). The above discussion shows that changing the line length L in fig. (12.7.4) is necessary to change the Trapatt frequency but that there has to be a tuner for the Impatt frequency so that the diode finds a resonator for a harmonic while the length L is again an odd multiple

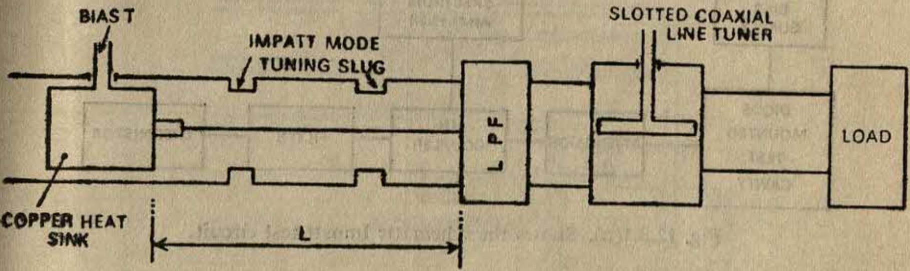


Fig. 12.7.4. Shows the TRAPATT oscillator circuit at MW frequencies.

of quater wavelength long for the Trapatt frequency when the reflected Trapatt harmonic voltage pulse may sustain itself to initiate at each Trapatt cycle. Normally Trapatt mode is not self-sustaining and it must be triggered for each cycle by a preceding Impatt oscillation. Thus we see that the recovery current is expected to have some control over the frequency as well since, if by a period of the Trapatt cycle, the diode has not recovered to its biased state the voltage will not be as effective in producing a fresh shock front as is necessary. Pulsed power at KW level at 1-2 GHz has been obtained with Trapatt oscillators. They find application in phased array Radar system where the MW beam is electronically scanned by controlling the excitation phase of each element of the array.

12.8 Impatt Mounting and Testing

The diode is mounted in a MW circuit and connected to a load through a matching network. An experimental schematic circuit for testing high efficiency Impatt oscillators is shown in fig. (12.8.1a). In this circuit the diode mounted cavity can be either a co-axial mount as shown in fig. (12.8.1b) or a reduced height waveguide as shown in fig. (12.8.1c). The reduced height-guide is used to match the diode negative resistance (which is small) to the load. Sometimes the matching function is accomplished with a raised cosine taper to reduced height-guide as shown in fig. (12.8.1d). In this case note that the diode is resonated with a sliding short. The d.c. bias is fed to the diode through a bias T which presents a short circuit looking back to the bias supply. To remove heat from the device, the device is always inserted through a massive heat sink. In the above test

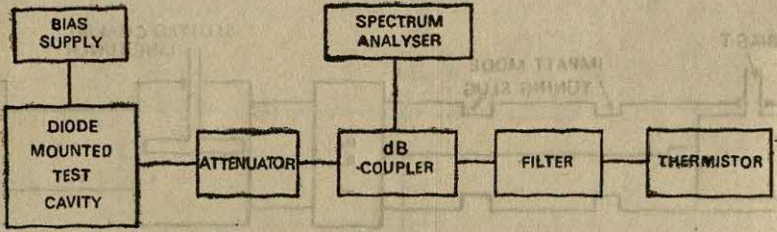


Fig. 12.8.1(a). Shows the schematic Impatt test circuit.

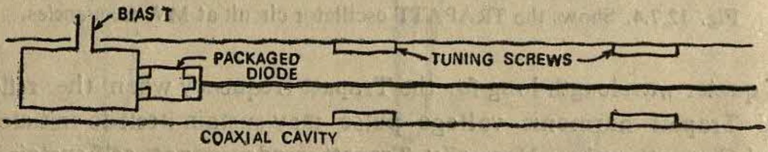


Fig. 12.8.1(b). Shows the diode mounted in a co-axial cavity.

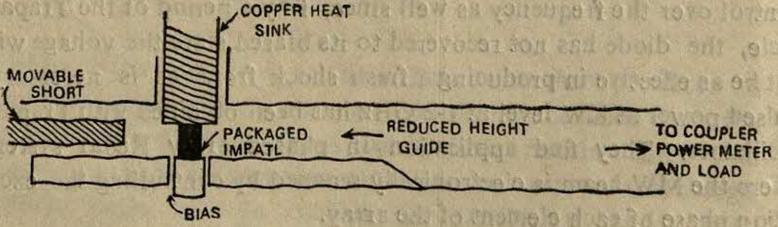


Fig. 12.8.1(c). Shows the reduced height wave guide with mounted diode.

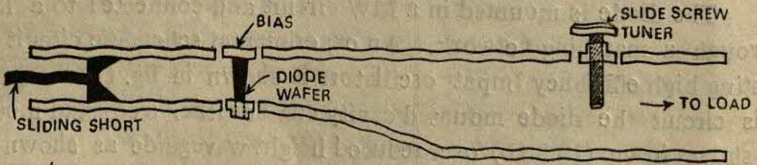


Fig. 12.8.1(d). Shows the Raised Cosine Taper Guide. Copyright © 1966 IEEE Trans ED-13 B.C. De Loch Jr. and R.L. Johnston by permission. Ref. 7.

circuit only about 10% MW power goes through the coupler to the spectrum analyser and rest to the detector. Suitable attenuators are incorporated to prevent over loading the power measuring detector and the filter is used to eliminate spurious signals far from the oscillation frequency. The wave-guide cavity possesses a higher Q factor than a co-axial system. This leads to a greater stability and lower FM noise in wave-guide circuits. Frequency modulation can be performed by modulating the bias current or by electronically varying the circuit susceptance using a reverse biased varactor diode in the cavity as shown in fig. (12.8.1*e*). Higher Q limits the ability to electronically tune for frequency modulation of the oscillator. The FM noise arises from jitter from cycle to cycle. The large amplitude R.F. voltage results in the creation of negative resistance in the diode over a much wider frequency spectrum. This requires a careful design of the out-of-band circuit characteristics, particularly in the bias connection. Proper tuning of the harmonic components of the waveform has a significant effect on oscillator efficiency. The co-axial circuit is more useful because it is relatively easy to model for analytical studies. Fig. (12.8.2) shows a circuit designed for single frequency operation. The equivalent circuit is also shown in the inset. The diode is

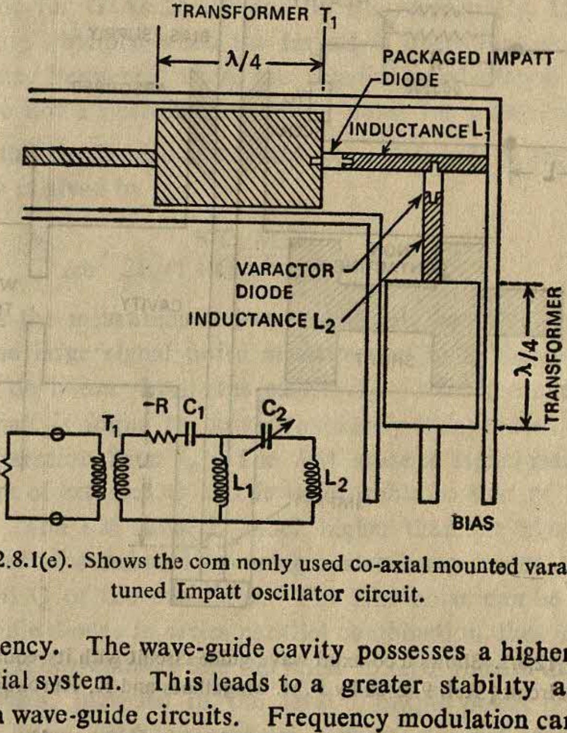


Fig. 12.8.1(e). Shows the commonly used co-axial mounted varactor tuned Impatt oscillator circuit.

mounted at the end of a co-axial transmission line which is terminated at the other end by a matched load. A low impedance matching section $\lambda/4$ long is inserted adjacent to the diode. The co-axial line

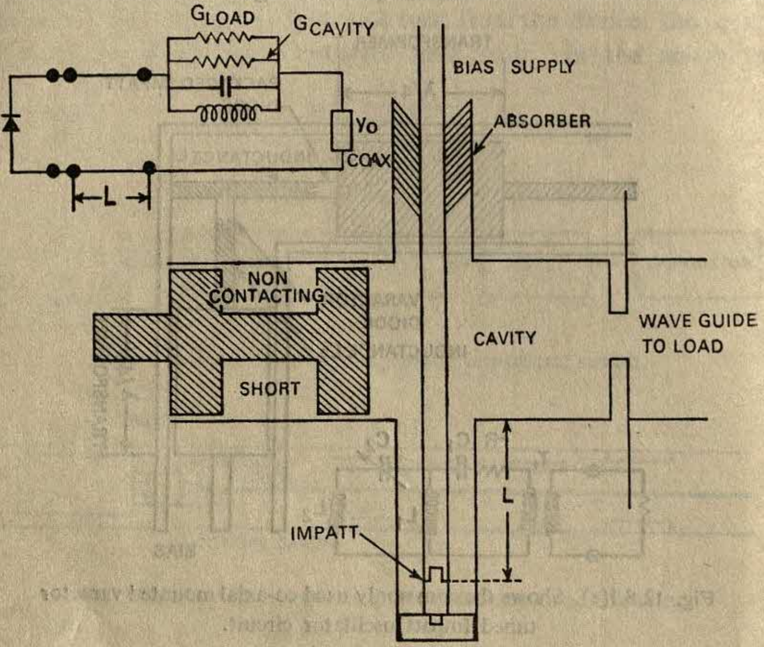


Fig. 12.8.2. Shows a co-axial wave guide circuit with its equivalent circuit [Ref. 9, used by F.M. Magalhaes and K. Kurokawa].

is series loaded with a high Q wave-guide cavity at a distance L from the diode as shown in the figure. In this circuit the sliding short is adjusted to optimise coupling from the co-axial line to the wave-guide cavity. The coupling between the cavity and the wave-guide system is adjusted by rotation of the cavity at the swival joint.

12.9 Noise

Impatt diodes working depends on avalanche ionisation process and on the transit time hence the noise-generating mechanisms are significant. The oscillator FM noise measure as given by Josenhaus (cf. ref. 5, 8) for small signal case as

$$M = \frac{P}{B}\left[\frac{Q_{ex}\,\Delta f_{rms}}{f_0}\right]^2 / k\,T_0 \qquad \cdots \quad (12.9.1)$$

where Q_{ext} the external circuit Q, B the measurement band width, f_{rms} the measured *rms* frequency deviation and P the oscillator power output at oscillator frequency f_0. Typical values of M in Si are, for small signal 40 *d*B and for large signal 55 *d*B, when working near the optimum efficiency. For Ge these are 30 *d*B and 40 *d*B while for GaAs 25 *d*B and 35 *d*B respectively. Lower noise measures are obtained when the Impatt is operating at twice the transit time frequency where the negative conductance is low and is therefore not a desirable operating point for many cases.

For the single side band AM noise (cf. ref. 5) the noise to signal ratio is given by

$$\left.\frac{N}{C}\right|_{ssB} = \frac{KT_0\, M\, B}{2P_0(1 + Q_{ext}^2\, f_m^2/f_0^2)} \qquad \cdots \quad (12.9.2)$$

where B is the measurement band width, P_0 oscillator power output, M the large signal noise measure and f_m is the measurement frequency deviation from the carrier f_0. Near the carrier, noise in Si—Impatt is found to be approximately independent of the frequency separation from f_0. The AM noise to signal ratio is nearly independent of external Q and is comparable to that of a Klystron oscillator. The FM noise is much higher than the klystron. The FM noise can be reduced by using a stabilising cavity that raises the external Q of the oscillator. The AM noise can be reduced by using multiple diodes in series parallel combination, thus making the R.F. voltages correlated while the noise voltages remain uncorrelated. In both cases, increase in the level of output power decreases the noise so long field suppression effects do not dominate the performance. $(1/f)$ noise do not occur from avalanche but may arise from upconverted flicker or $(1/f)$ trapping noise.

12.10 Baritt Diode

In the Impatt diode the avalanche multiplication is the cause of particle current pulse formation. Instead, forward biased injection of minority ccrriers in a *p-n* junction diode over the junction barrier can be used as the source of particle current pulse. This particle current pulse can be led through a swept out drift region in diodes, like P^+-N-I-P^+, P^+-N-P^+ or P^+-N-P-P^+ when it would acquire delay due to transit time through the drift space. To the fundamental component of the induced current the diode presents a negative resistance. These diodes are called BARITT (BARrier Injection Transit Time) diodes.

The diodes mentioned above can be thought of as swept out floating base transistor where emitter to collector voltage is applied. Hence the field in the reverse biased region in general, is not large everywhere. This gives the possibility of working at low voltages like 7 volts at 10 GHz when an Impatt diode in a corresponding situation requires at least a voltage drop of 50 volts in its drift region. This low voltage operation is however associated with reduced power output. Nevertheless, these may be useful for

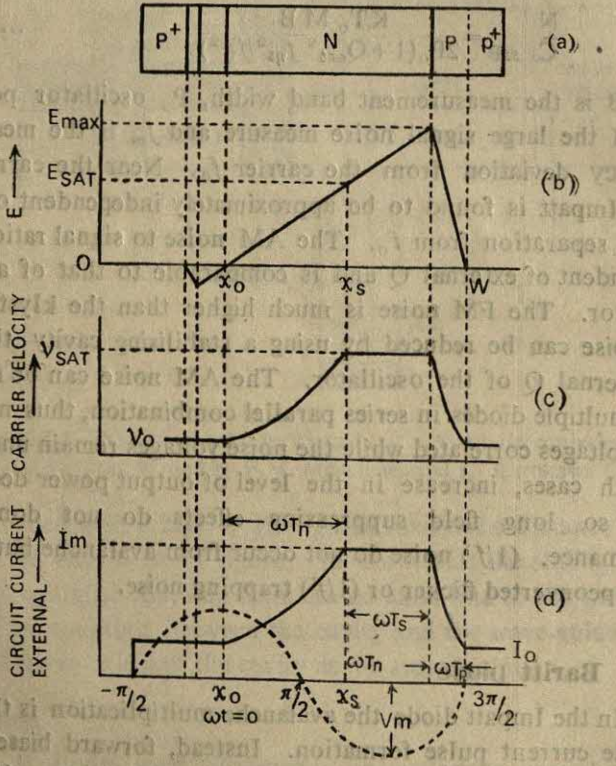


Fig. 12.10.1. (*a*) shows the swept out P+-N-P-P+ diode ; (*b*) shows the electric field ; (*c*) shows the carrier velocity ; and (*d*) The external circuit current as related to the R. F. voltage across the diode. (By permission Ref. 11. Copyright © 1977 IEEE Trans. ED 24 Authors : H. N. NGUYEN & G. T. HADDAD)

applications like doppler radar system where low d.c. power consumption with low radiated power is important.

The field, carrier velocity, induced current and R. F. voltage waveforms as expected in a P^+-N-P-P^+ diode is shown in fig. (12.10.1a—d). It is assumed that the particle current pulse gets injected at the peak of the RF voltage. The carriers are injected across the forward biased junction. At x_0 onwards these carriers are aided by the electric field. At $t=0$ and $x=x_0$ the carriers had a diffusion velocity v_0 that increases as time passes due to the field as

$$(dx/dt)=v(x)=\mu\,E(x)=\mu\,qN\,x/\varepsilon\varepsilon_0 \quad \cdots\cdots \quad (12.10.1)$$

till the field at $x=x_s$ reaches the velocity saturation field E_s. The transit time from x_0 to x_s is assumed T_n. So long the field is above E_s that is required for carrier velocity saturation, maximum induced current I_m is obtained. The transit time through this region is assumed as T_s. Thereafter over a period of time T_p the carrier velocity falls from the saturation value to the diffusion velocity, correspondingly the induced current falls to the cut-off bias value, I_0. As an optimistic diagram the R F. voltage is drawn with the same time-scale. The current I_0 is low enough such that the d.c. power loss during regions like $\omega t=-\pi/2$ to zero can be ignored, also the R. F. output in these regions is neglected.

From equation (12.10.1) we have

$$v=v_0\exp(\mu\,qNt/\varepsilon\varepsilon_0) \quad \cdots \quad (12.10.2)$$

where $v_0=(\mu q\,Nx_0/\varepsilon\varepsilon_0)$ ··· (12.10.3)

So, the induced current increases till $t=T_n$ as

$$I=I_0\exp(\mu q\,Nt/\varepsilon\varepsilon_0) \quad \cdots \quad (12.10.4)$$

Between $t=T_n$ to $t=T_n+T_s$ the induced current

$$I=I_m \quad \cdots \quad (12.10.5)$$

Thereafter the induced current falls due to decreasing field (proportional to distance) hence exponentially again to I_0 and can be written as

$$I=I_0\exp qN\,\mu\,N_a(3\pi/2\omega-t)/\varepsilon\varepsilon_0 \quad \cdots \quad (12.10.6)$$

between values of $\omega t=(3\pi/2-\omega T_p)$ to $(3\pi/2)$ where N_a is the p layer doping density. From equations (12.10.2) to (12.10.6) we define

$$a = (\mu q\, \mathrm{N}_\omega\, \mathrm{T}_n/\omega\varepsilon\varepsilon_0) = (\mu q\, \mathrm{N}_{a\omega}\, \mathrm{T}_p/\omega\varepsilon\varepsilon_0) = \log b$$
$$= \log (v_s/v_0) = \log (x_s/x_0) = \log (\mathrm{I}_m/\mathrm{I}_0) \qquad \cdots \quad (12.10.7)$$

The R.F. output can be obtained by adding the contributions from the three sections of the induced current as defined in fig. (12.10.1) assuming the peak R.F. voltage to be V_m as

$$\mathrm{P}_0 = \int_0^{3\pi/2} \mathrm{I}\, \mathrm{V}_m \cos \omega t\; d(\omega t) \qquad \cdots \quad (12.10.8)$$

while the d.c. current can be found by

$$\mathrm{I}_{dc} = \left(\frac{1}{2\pi} \int_0^{2\pi} \mathrm{I}\, d\theta\right) \qquad \cdots \quad (12.10.9)$$

where the contribution to the integral of I_0 for the region $\omega t = -\pi/2$ to zero can be ignored. The efficiency can be calculated from $\eta = \mathrm{P}_0/\mathrm{V}_{dc}\, \mathrm{I}_{dc}$ and is

$$\eta = \frac{\mathrm{V}_m}{\mathrm{V}_{dc}} [\{a\omega \mathrm{T}_n (b \cos \omega \mathrm{T}_n - 1)$$
$$+ b\omega^2 \mathrm{T}_n^2 \sin \omega \mathrm{T}_n\}/b(a^2 + \omega^2 \mathrm{T}_n^2)$$
$$- \{ba\omega \mathrm{T}_p \sin \omega \mathrm{T}_p + \omega^2 \mathrm{T}_p^2 (1 - b \cos \omega \mathrm{T}_p\}/b(a^2 + \omega^2 \mathrm{T}_p^2)$$
$$+ \{\sin \omega (\mathrm{T}_n + \mathrm{T}_s) - \sin \omega \mathrm{T}_n\}] \Big/ \left[\frac{\omega}{ab} (\mathrm{T}_n + \mathrm{T}_p)(b - 1) + \omega \mathrm{T}_s\right] \qquad \cdots \quad (12.10.10)$$

The computer calculations of the conversion efficiency indicate that the efficiency is maximum for $\mathrm{T}_s = 0$, $\omega \mathrm{T}_n = \pi$ and $2\omega \mathrm{T}_p = \pi$. For this situation, the efficiency in terms of $(\mathrm{V}_m/\mathrm{V}_{dc})$ can be written as

$$\eta = \frac{\mathrm{V}_m}{\mathrm{V}_{dc}} \frac{2a}{3(1-b)} \left[\frac{a(1+b)}{a^2+\pi^2} + \frac{(\pi + 2ba)}{4a^2+\pi^2}\right] \qquad \cdots \quad (12.10.11)$$

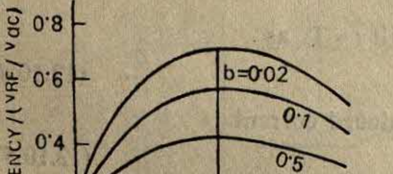


Fig. 12.10.2. Shows the efficiency in terms of the ratio $\left(\frac{\mathrm{V}_m}{\mathrm{V}_{d.c}}\right)$ for $= b\left(\frac{\mathrm{I}_m}{\mathrm{I}_0}\right)$ as a parameter for $\mathrm{P}^+ - \mathrm{N} - \mathrm{P}^+$ diode.

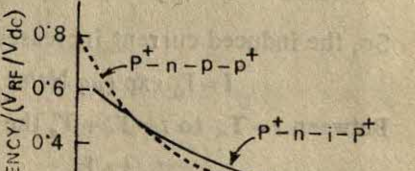


Fig. 12.10.3. Shows the efficiency of two types of diodes.

For $T_p = 0$ this diode reduces to P^+-N-I – P or P^+-N-P diode when at $\omega t = 3\pi/2$ the induced current drops from I_m to zero. The efficiency for these cases as function of ωT_n and b are shown in figs. (12.10.2) and (12.10.3) (cf. ref. 11). Extensive computer calculations (11) conclude that the P^+-N-P-P^+ gives the best efficiency and has a low d.c. power requirement while the P^+-N-I-P would give the best power performance. N-type GaAs would give the best overall performance with N-substrate Si being close to it. Practical Baritt diode has been seen to work at 6·5 GHz at $P_0 = 100\ m$W and 2 percent efficiency with 15 dB noise figure (cf. ref. 11).

References

1. Negative Resistance Arising From Transit Time in Semiconductor Diodes, W. Shockley, Bell System Tech. Jour. 33, (1954), p. 799.
2. A Proposed High-Frequency Negative Resistance Diode, W.T. Read, Bell System Tech. Jour., 47, (1964), p. 369.
3. Physics of Semiconductor Devices, S.M. Sze, John Wiley and Sons, 1969, p. 200.
4. Large signal Analysis of a Silicon Read Diode Oscillator, D.L. Scharfetter and H.K. Gummel, IEEE Trans. ED. Vol. ED.-16, No. 1. p. 64—67, 1969.
5. Topics on Solid State and Quantum Electronics, W.D. Harshberger, John Wiley and Sons, Inc., (1972).
6. The Effect of Unswept Epitaxial Material on the Microwave Efficiency of IMPATT Diodes, S.R. Kovel and G. Gibbons, Proc. IEEE Vol. 55, 2066, 1967.
7. Avalanche Transit Time Microwave Oscillators and Amplifiers, B.C. Deloach, Jr. and R.L. Johnston, IEEE Trans. ED-13 181—186, Jan. 1966.
8. Semiconductor Devices and Integrated Electronics by A.G. Milnes, Van Nostrand Reinhold Company, 1980.
9. A Single-tuned Oscillator for Impatt Characterisation F.M. Magalhaes and K. Kurokawa. Proc. IEEE (correspondence), 58, p. 831-832, 1970 May.

10. A Low Noise Metal-Semiconductor-Metal (MSM) Microwave Oscillator. Bell Syst. Technical Journal, Vol. 50, p. 1695, 1971.
11. Effects of Doping Profile on the Performance of Baritt Devices, Hien Nguyen-BA and George I. Hadded. IEEE Trans. ED 24, No. 9, 1977.

Problems

1. Show that for a Read diode the anti-resonant frequency varies as the square root of the current density and current variation causes variation of the frequency at which maximum negative conductance occurs.
2. Show that the separation between the anti-resonant frequency and the frequency at which maximum negative conductance occurs, decreases as the current increases.
   Hint : Start from equation (12.3.25), find $[d(\theta_m - \theta_r)/d\theta_m]$ to show it is negative.
3. Show that the Read diode admittance magnitude increases with current.
   Show that the maximum of this quantity occurs at $f = 3.f_r$.
4. Plot G vrs. B curves for a silicon Read diode of drift region width $d = 10$ microns. Assume $v_s = 10^7$ cm/sec, $\varepsilon = 12$, $x_a = 0{\cdot}4$ microns and $\overline{\alpha'_0} = {\cdot}02$ per volt. Calculate spot values for 50, 100, 250, 400 amp/cm^2 at 8 GHz, and at 250 amp/cm^2 for 3, 5, 7, 8 and 10 GHz.
5. Plot the variation of intrinsic Q of the Read diode of problem (4) above at the specified points.
6. Find how does the minimum Q of a Read diode vary with current.
7. Argue to establish that current variation can give frequency tuning but is a very unsafe method.

Chapter 13

GUNN DEVICES

In this Chapter we shall consider bulk negative differential resistance devices and their uses. In 1961 Ridley and Watkins (ref. 1) and independently in 1962 Hilsum (ref. 2) proposed that bulk differential negative resistance effects are obtainable from the field induced transfer of conduction electrons from a low energy high mobility valley to higher energy low mobility satellite valleys in materials like GaAs which possess such suitable band properties (cf. fig. 2.7.1). In 1963 Gunn (ref. 3) observed that when the applied d.c. electric field across an N-type sample of GaAs or InP exceeded a critical threshold value of a few kilovolts per centimeter oscillations at microwave frequency were obtained. Later, Kroemer (ref. 4, 5) pointed out that these oscillations were consistent with the Ridley, Watkins and Hilsum (RWM) proposed mechanism. Experiments on GaAs subjected to pressure and the experiments on alloys like $GaAs_{(1-x)}P_x$ showed that the threshold field decreases with decreasing energy difference between the valley minima. This convincingly established that the Gunn oscillations were due to RWH mechanism mentioned above. Later, it has been found that (ref. 2) materials like CdTe, ZnSe and InAs under uniaxial pressure, also show RWH mechanism at high fields. We shall, however, confine our discussion of this effect with reference to the material N-type GaAs.

13.1 Velocity Field Characteristics

A typical property of many direct band gap semiconductors is (cf. fig. 2.7.1) that the effective mass of the central valley is considerably lower than the free electron mass. In GaAs $m_c = 0{\cdot}07 m_o$ (cf. ref. 6), this leads to high electron mobilities despite the existence of strong electron scattering inside the semiconductor (the so-called intervalley scattering). In GaAs this mobility is 8000 cm^2/v. sec. at room temperature. On the other hand, the effective masses in satellite valleys are very much larger, on the average, of the order of free electron mass. The lowest set of satellite valleys is about 0·36 ev. above the central valley. This energy is very much larger than $kT = 0{\cdot}025$ ev. Hence at room temperature and for low electric fields, nearly all electrons will be located near the bottom of the central valley where they have high electron mobility. A suitably strong electric field is capable of accelerating the electrons to rather high energies of the order of satellite valley energy. When this happens, the electrons are scattered into the satellite valleys where they have a much higher effective mass and hence a much lower mobility. So, they contribute much less to the current. This current reduction occurs rather abruptly because of the presence of strong intervalley scattering between the different satellite valleys. As a result, the overall current drops, and the material exhibits a bulk negative differential conductivity.

Ruch and Kino (ref. 7) measured the velocity field characteristics of GaAs up to about 14 kv/cm, which agrees well with the theoretically predicted model. In their measurement, a high resistivity 10^6—10^8 Ohm-cm GaAs was provided with a gold contact cathode on top of SiO and an Ohmic Sn anode. The SiO layer was used to prevent hole injection. An electron beam was swept past a slit that allows a short pulse (of the order of 0·1 nano sec.) to be incident on the gold cathode. When the electron pulse hits the cathode a bunch of charge of electrons is formed inside the semiconductor.

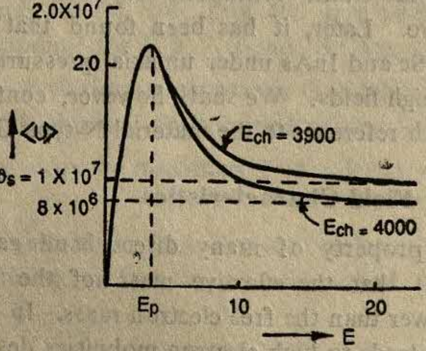


Fig. 13 1.1. Velocity field characteristics of GaAs from results of different workers [Ref. 5, H. Kroemer, Copyright : John Wiley & Sons, INC 1972, Copyright © 1968, IEEE, Trans Ed-15]

This bunch induces a current in the terminals of the sample during its transit time. To avoid heating, a pulsed electric field was applied to accelerate this bunch of electron. As the meterial has high resistivity, the field is primarily uniform. Thus from a measurement of the width of the induced current pulse in time scale and from the known length of the sample, the drift velocity was determinded accurately. The field corresponding to this velocity is known from a knowledge of the applied pulse voltage and the length of the sample. Diffusion only affects the rise and fall times of the output pulse but not its duration.

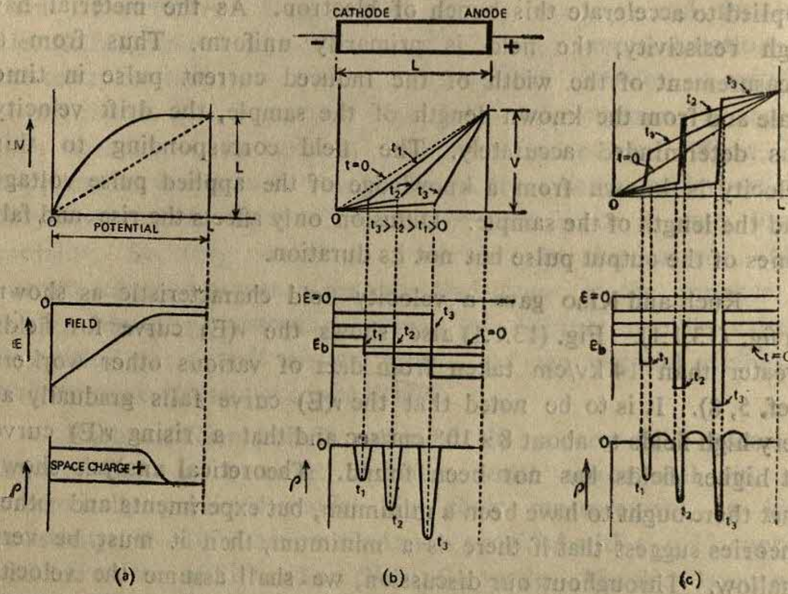
Ruch and Kino gave a velocity field characteristic as shown in fig. (13.1.1). Fig. (13.1.1) also shows the $v(E)$ curve for fields greater than 14 kv/cm taken from data of various other workers (ref. 5, 8). It is to be noted that the $v(E)$ curve falls gradually at very high fields to about 8×10^6 cm/sec and that a rising $v(E)$ curve at higher fields has not been found. Theoretical analysis shows that there ought to have been a minimum, but experiments and other theories suggest that if there is a minimum, then it must be very shallow. Throughout our discussion, we shall assume the velocity field characteristic as shown in fig. (13.1.1) i.e., at very high fields the valley velocity is truly constant and is equal to v_s. The $v(E)$ curve over the entire range of field can be well approximated by

$$v(E) = [\mu_0 E + v_s(E/E_{ch})^k]/[1 + (E/E_{ch})^k] \quad \cdots \quad (13.1.1)$$

where the low field mobility $\mu_0 = 8000$ cm^2/v sec., $E_{ch} = 4000$ v/cm, $v_s = 8 \times 10^6$ cm/sec and $k = 4$. If one takes $k = 4$, $E_{ch} = 3{\cdot}9$ kv/cm and $v_s = 10^7$ cm/sec, eqn..(13·1) gives excellent fit to Ruch and Kino's results (cf. ref, 5).

13.2 Current Voltage Characteristics

The d.c. current voltage characteristics of the transferred electron devices (TED) (made up of two ohmic contacts at the two ends of an N-type GaAs crystal of dimension of the order of 100 microns), as reported in the early sixties were much conflicting. In that it was observed that the current follows Ohms law upto a certain field strength but many workers reported (cf. ref. 6) a fall of current at high voltages while some others reported and justified a monotonic increase. The reasons will be well understood when the transient behaviour of the diode has been discussed. Hence we discuss the transient behaviour of the diode at the outset.



Fig, 13.2 1. Shows the potential, field and space charge for three different idealised distinct charge [Ref. 4, 5, Copyright : John Wiley & Sons, INC. 1972]

During the period between 1964 to 1970 (cf. ref. 5, 6) various workers reported capacitance probe measurements on the transient field distribution in diodes at fields higher than the critical field. All these indicated that if the doping is high and sample length large (then the common features of these reports showed that), three typical idealised field patterns could be formulated in simple terms, as shown in fig. (13.2.1). These can be grouped under the following assumptions :

(1) There is a transient repetitive charge movement in the diode, (2) the charge at any location can be accounted for by the difference of the field on either side of it, (3) the field is constant on either side of the charge.

The field types are discussed below. Under constant applied voltage conditions at fields greater than the critical field E_p (cf. fig. 13.1.1), the potential and field distributions are as shown in fig. (13.2.1). The understanding of this figure is important as it leads to an understanding of the constant applied voltage diode transients.

(*a*) *The Static Case* :

In the first case fig. (13.2.1a) shows that, a static potential drop has occurred at the cathode contact at constant applied voltage and even if the field given by $E_b = (V/L)$ is greater than E_p, it is small in the rest of the sample.

(*b*) *The Moving Accumulation Layer Mode* :

In this a growing negative bunch of charge moves from cathode to anode repetitively, as one bunch terminates on the anode, another starts at the cathode.

(*c*) *The Moving Domain Mode* :

In this a growing domain (a positive charge followed by a negative growing charge) moves repetitively from the cathode to the anode, as one domain terminates on the anode another starts from the cathode.

For an understanding of these constant applied voltage transients, we concentrate on the second case. Fig. (13.2.1b) shows that although the applied voltage produces average field $(V\ L) > E_p$, as time passes the field at the cathode becomes gradually lower than $E_b = \frac{V}{L}$ and the field at the anode rises to make up for the total voltage. The field change is sharply located as shown in fig. (13.2.1b) and increases as the charge bunch travels towards the anode. This implies that the charge (dE/dx) is negative, hence an accumulation layer of charge travelling towards the anode. In the next article it is shown that this type of charge movement can be repetitive for constant voltage bias exceeding the critical value $E_p.L$.

In contrast, the first fig. (13.2.1a) shows that a static charge distribution has occurred, most of the applied voltage has dropped at the cathode region and near the cathode there is a stationary positive charge attached to the cathode. This has led to a large field at the cathode region while the field is low beyond this fixed positive charge on the anode side.

In the third type in fig. (13.2.1c) it is seen that under the same applied voltage, the fields on both the cathode and anode sides of the change are falling with time while the field within the charged region increasing. This implies that a depletion region followed by a growing accumulation region (a domain) is on transit. In this case, when the depletion region has been delivered to the anode, the accumulation region remains to be delivered ; the situation is like fig. (13.2.1b), when the growing accumulation layer is being delivered to the anode. Moreover, the field patterns, with it the charges, depicted in figs. (13.2.1b) and in (13.2.1c) are repetitive in nature. The repetition time is of the order of the transit time of the carriers.

13.3 Time Evolution of Accumulation Charge

In order to understand the time evolution of the accumulation layer (AL) charge, we here follow the lines indicated by Kroemer and assume (cf. ref. 5) that a certain voltage $V = E_p.L = E_b.L$ exists across the diode with ohmic contacts on both ends. A step voltage δV has been applied thus increasing the voltage to $V + \delta V$. This δV voltage step causes a displacement charge injection at the cathode end of value $Q_o = \varepsilon_o \varepsilon \, \delta V/L$. Remembering the assumptions mentioned above in 1—3, assume that the field on the cathode side of the charge is $(E_b + E^-)$ while the field on the anode side of the charge is $(E_b + E^+)$. The difference δE of these two fields supports the charge Q, hence

$$(E^+ - E^-) = \frac{Q}{\varepsilon \varepsilon_o} \quad \cdots \quad (13.3.1)$$

If the currents leaving and entering the charge regions are J^- and J^+ respectively, we get

$$\delta J = (J^+ - J^-) = qN \left(\frac{\delta v}{\delta E}\right) \delta E \quad \cdots \quad (13.3.2)$$

The rate of growth of the charge is given by $(\delta Q = -\delta J.\delta t)$ which may be written as

$$\frac{\delta Q}{\delta t} = -qN\,(\delta v/\delta E).Q/\varepsilon\varepsilon_o \quad \cdots \quad (13.3.3)$$

which indicates that as soon as the δV voltage step has been applied (while increasing the voltage) a bunch of electron charge $Q_o = \varepsilon\varepsilon_o \delta V/L$ is introduced into the sample at the cathode. This charge, as it travels with an average velocity v_o, changes with time as

$$Q = Q_0 \exp(-t/T_d) \quad \cdots \quad (13.3.4)$$

where $T_d = \left(\varepsilon_0 \varepsilon / qN \dfrac{dv}{dE}\right)$.

We speak of such an average because the maximum and minimum values of velocity differ by a factor of two only and so we think of the main bulk of charge to move with this average, while there can be local changes. So, we can speak of a transit time given by $T_t = (L/v_0)$. Thus, we see that if (dv/dE) is negative the charge, as given by (13.3.4), grows as it travels. The overall growth of the charge that occurs in one transit time is equal to

$$Q(T_t) = \frac{\varepsilon\varepsilon_0\, \delta V}{L} \exp\left(\frac{L}{v_0 T_d}\right) = \frac{\varepsilon\varepsilon_0\, \delta V}{L} \exp\frac{NL}{M} \quad \cdots \quad (13.3.5)$$

where $M = \dfrac{\varepsilon\varepsilon_0 v_0}{q\,|dv/dE|}$ $\cdots$ (13.3.6)

For GaAs, taking typical values of $v_0 = 1{\cdot}6 \times 10^7$ cm/sec,

$\varepsilon_0\varepsilon = 8{\cdot}85 \times 10^{-14} \times 10{\cdot}9$ F/cm and $(dv/dE) = \mu = -8000/3$,

eqn. (13.3.6) gives $M = 0{\cdot}4 \times 10^{11}$.

Thus, we see that the primary accumulation layer charge injected by voltage increase δV can grow by a maximum amount of $\exp\frac{NL}{M}$. That is, the growth is dependent on the NL *i.e.*, doping level and length product as also on M, that is, average drift velocity and the magnitude of negative differential mobility (thus on the material as well). If $NL = M$, the charge growth is only by a factor e while $NL \gg M$ implies large charge growth and $NL < M$ indicates no charge growth practically. We also note that $NL > M$ indicates that the accumulation charge grows, but does not disperse as would be expected in a positive mobility material or for low fields in GaAs. That is, the charge is held together by negative differential mobility which effectively is working to modify coulomb repulsion to an effective attraction amongst the charge particles. In the above discussion, diffusion has not been taken into account. But diffusion would only cause spreading of the charge, hence would act against this attractive force and the shape of the accumulation layer would be determined by the competition between the two mechanisms.

Next we examine the time evolution of the cathode field. When the charge is at a distance x from the cathode, we have

$$(E_b + E^-)x + (E_b + E^+)(L - x) = \delta V + V$$

where $E_b x + E_b(L - x) = V$

hence $E^- x + E^+(L - \lambda) = \delta V \qquad \cdots \quad (13.3.7)$

From equation (13.3.7), (13.3.1) with (13.3.4) we have

$$E^- x + \left(E^- + \frac{Q}{\varepsilon \varepsilon_0}\right)(L - x) = \delta V$$

or $$E^-.L + (L - x)\frac{\delta V}{L} \exp(\nu_0 t/\nu_0 T_d) = \delta V$$

or $$E^- = \frac{\delta V}{L}\left[1 - \frac{L - x}{L} \exp\left(\frac{Nx}{M}\right)\right] \qquad \cdots \quad (13.3.8)$$

Substituting $N/M = \theta/L$ equation (13.3.8) can be written as

$$E^- = \frac{\delta V}{L}\left(1 - \frac{L - x}{L} \exp \frac{\theta x}{L}\right) \qquad \cdots \quad (13.3.9)$$

For different values of $\frac{NL}{M} = \theta$, the cathode side field can be drawn as a function of (x/L) as shown in fig. (13.3.1). It is to be noted that—

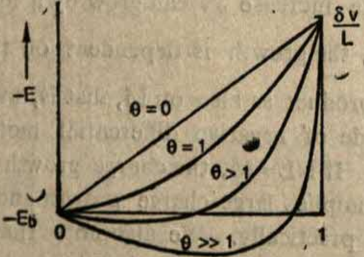


Fig. 13.3.1. Shows the variation of the field at cathode side of the accumulation layer space charge. Note that for $\theta > 1$ the total cathode field initially decreases in magnitude.

(1) for $\theta = \frac{NL}{M} = 0$, the expression (13.3.5) indicates that the charge growth is nil while the cathode side field increases linearly with the location of the charge.

(2) for $\theta = \frac{NL}{M} = 1$, the growth factor is only e while the initial rate of cathode field change is zero as can be seen by differentiating (13.3.9) with respect to x that zero slope occurs at $(x/L) = (\theta - 1)/\theta$.

(3) For $\theta = NL/M > 1$, the cathode field initially decreases from the applied value $E_b = \frac{V}{L}$ and later rises to the value $(V + dV)/L$. Typically when $NL/M \gg 1$, say, for 9·2, the location of the zero of the field deviation from V/L is at $x/L = 0{\cdot}9999$. That is, for large NL/M values; for most of the time, the field at the cathode side of the charge is below the nominal value of $E_b = V/L$. Only when the charge has reached very near the anode, the cathode field rises to E_b and in a very short time to the value $(E_b + dV/L)$. It can be easily shown that the maximum deviation of the cathode field from E_b is $\frac{\delta V}{L\theta}$ [exp $(\theta - 1) - \theta$] which is of the order of 359 $(\delta V/L)$ for $\theta = 9{\cdot}2$. This change occurs in a time $(8{\cdot}2/9{\cdot}2)L/v_0$ while a field change of 360 $\frac{\delta V}{L}$ occurs in the subsequent time $(L/9{\cdot}2v_0)$. That is the cathode field increase as compared to the decrease is at least 8·2 times sharp for $NL/M = 9{\cdot}2$.

Following Kroemer, (cf. ref. 5) we introduce here the "reinjection feedback" concept. We have seen that as the primarily injected AL moves away from the cathode it grows for large NL/M values while the cathode field changes as well. This change of cathode field implies current injection changes with it, say, proportionately. For the case of $NL = M$, the cathode field increases slowly throughout the transit time of the charge, but the total amount of change is $(\delta V/L)$. Hence a charge equal to the primary AL's charge enters, but is distributed in time; hence in space; and so must be of lower total value, while the primary AL grows by a factor of e only. It is to be remembered that small values of distributed charges may be ignored in view of the presence of a time independent background charge. After the primary AL is delivered to the anode, useful part of the secondary charge that gets injected while the primary AL was on transit, (can be thought as located at the centre of this charge) in its turn, moves and grows by the same factor e and during this time a tertiary charge of still more distributed nature gets introduced by the time the secondary charge is delivered to the anode. Thus, we see that for values of $NL = M$, a voltage step δV causes a ringing of the device giving successive current pulses, with time gaps a little less than the transit time (L/v_0), that gradually decays down to zero.

If, of course, $NL \gg M$, the cathode field initially drops. By the above argument, it would seem that the current would decrease. But this decrease occurs over a length of time : hence when one considers the time independent background charge due to E_b this decrease may not be observable as was the small values of the distributed charges in case of $NL = M$. Moreover, in this case, the primary AL charge grows by a large factor. For the typical case of $NL/M = 9{\cdot}2$ by as much as 10^4 times. That is, a charge of $\frac{\varepsilon_0 . \varepsilon \, \delta V}{L} \times 10^4$ is delivered to the anode with a consequent rise of the cathode field by 350 $\delta V/L$ in time $(L/9{\cdot}2 \; v_0)$. This large change of cathode field, in such a short time, introduces a larger amount of charge at the cathode than the primary AL's introduced value and that it is localised in time of $(L/9{\cdot}2 \; v_0)$ in this particular example. This secondary AL moves and causes introduction of still larger tertiary current pulse in its turn. Thus for $NL \gg M = 10^{11}$, it is expected that the device will produce growth of oscillation which will ultimately be limited by the various losses such as diffusion, non-linearities and ohmic losses. Hence $NL \geqslant 10^{12}$ can be taken as the criterion for self-sustaining oscillations in the Gunn diode.

13.4 Imperfect Cathode

In the above discussion it was tacitly assumed that the cathode is well behaved in that equal change of cathode field produces equal change of charge injection or extraction. It was noted by various experimentalists (cf. ref. 6) that around any inhomogeneity in the sample an AL changes into a domain. This leads to turn our attention to the cathode, because it is the place of maximum inhomogeneity, for an explanation of the third type of observation mentioned in article (13.2). The explanation given below also follows the lines of reasoning given by Kroemer.

The hint lies in the point that, the current that the contact allows to flow as compared to the current the bulk can carry for a certain interface field between the contact region and the bulk. The bulk current versus the interfaces field characteristics may be obtained from the velocity field characteristics of the sample by multiplying with qN and is shown in fig. (13.4.1) plotted against the interface field. The contact or the contact characteristic against the interface field can then obviously be of only three types as shown in this figure. If the contact characteristic is one like the curve '*a*', then for any value of the interface field the contact can give more current than the bulk can carry. In such a case, an accumulation of electrons will always result at the interface and pure AL propagation from the contact is expected. If the contact characteristics is one like '*c*' then for a certain interface field, the contact gives less current than the bulk carries and at the interface a depletion of electrons would result. This depletion, in turn, will cause increase of the interface field and this is expected to continue until the interface field has been such as to establish a current equality of the bulk with the contact regions. In this case, therefore, a static potential distribution of the type as shown in fig. (13.2.1a) will result.

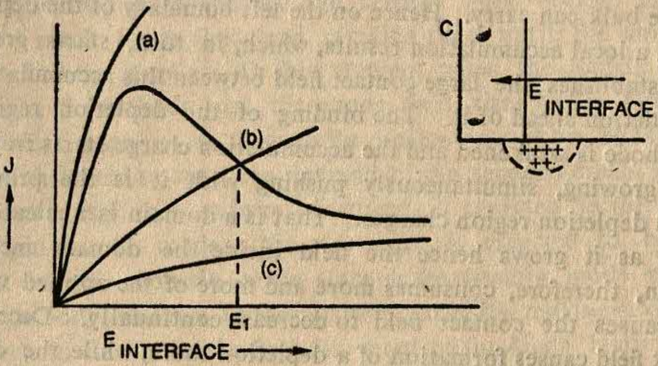


Fig. 13.4.1. The bulk current and cathode contact current as functions of the field at the interface between cathode region and bulk region. [Ref. 5, H. Kroemer, Copyright : John Wiley & Sons., INC. 1972. Copyright © Nov. 1968, IEEE Trans. Ed-15]

If it so happens that the contact characteristic is one like the curve '*b*', then for an interface field lower than E_1 (where bulk current equals the contact current) the contact gives less current than the bulk can carry and hence a depletion results at the interface. This depletion, in turn, increases this interface field until it becomes E_1. That is for such a contact characteristic, we have always a depletion region formed at the cathode contact bulk interface. If any thermal fluctuation decreases this interface field, the interface field allows more bulk current than that reaches there from the contact and hence the depletion enhances to restore the interface field to E_1. The situation is vastly different if the interface field momentarily becomes more than E_1, since in that case, the contact supplies more carriers than the bulk can carry. Hence on the left boundary of the depletion region, a local accumulation results, which, in turn, starts growing. This establishes the large contact field between this accumulation to the depletion ahead of it. The binding of the depletion region to the cathode is weakened and the accumulation charge starts travelling while growing, simultaneously pushing with it is the previously formed depletion region charges. That is a domain is nucleated. It travels as it grows hence the field inside the domain increases. Domain, therefore, consumes more and more of the applied voltage. This causes the contact field to decrease continually. Decrease of contact field causes formation of a depletion there, while the domain is in transit in which the accumulation layer grows continually. Thus the field within the domain grows until the accumulated charge becomes equal to the depletion charge when we speak of a matured or a stable domain. This stable domain obviously implies that the velocity within the domain has become equal to the velocities just outside so that no extra charge is entering or leaving the domain as shown in fig. (13.4.2) when diffusion is ignored. When the domain has matured, large voltage drop exists in the domain and very little voltage is available outside, resulting in small fields outside the domain on either side. This causes the cathode depletion to grow until the domain is matured. When the depletion region of the domain arrives at the anode, the situation is similar to the AL mode just before the AL delivery to the anode, except for the small cathode voltage rise due to the already formed depletion there. The domain mode is an oscillatory phenomena ; as soon as the AL is delivered to the anode, the cathode field rises to the bias value E_b causing electron injection whose growth and movement pushes the depletion with it and a new domain starts.

The time evolution of the field in both the AL and the domain mode has already been shown in article (13.2). It is to be noted that in the AL mode the anode field is large for most of the time except when the AL has just been delivered to the anode, when this field drops to the bias value momentarily. This causes the velocity at the anode to come near the peak velocity momentarily, then it decreases again as the anode field increases in the next cycle.

In the domain mode once the domain has started on its transit the field at the anode is lower than the bias value. As the domain grows this field goes through the critical peak velocity field E_p, attains a much lower field and remains there until the domain is delivered to the anode. At this moment, when the domain reaches the anode, the cathode and the anode field rises to the bias value through E_p (hence passes through the peak velocity) but to fall again through this field to meet the demand of the second domain growth. So, a more spiked waveform of current is expected in the domain mode than in AL mode. The AL mode is expected to be more efficient than the domain mode since in the domain mode most of the diode body remains in the positive mobility region as opposed to the AL mode where the field for most of the diode body remains in the negative differential mobility region for bias voltages well above E_p.L.

The speed with which domain maturity would occur is expected to depend on the NL product. For NL products greater than 10^{12}, the domain growth is seen to be complete within 20% of the transit time. Whether a domain or an AL would form depends on the cathode contact region's sample inhomogeneity.

13.5 Butcher's Equal Area Rule *(ref. 9)*

We have noticed that in the stable domain the field inside the domain is high, say, the highest value is E_m and the outside field low, assume it is E_0. In order to understand the extent of the stable domain or the values of the fields therein, we here find out the rule that governs the stable domain which is known as Butcher's equal area rule. For this, we consider the expression for the total current assuming for simplicity that the diffusion constant is field independent and the current is referred to the particle direction. We have,

$$J = (qn\, v(E) - qD \frac{\delta n}{\delta x} + \varepsilon_0 \varepsilon \frac{\delta E}{\delta t} \qquad \cdots \quad (13.5.1)$$

where in respect of fig. (13.2.1) the field is directed to the left, v is directed to the right ; hence **J** is directed to the right.

Since we know that the field pattern is moving, let us transform equation (13.5.1) in the moving reference frame. The reference frame is assumed to be moving with velocity u. In this frame the distance z at time t' is defined by

$$z = x - \int_0^t u(t)\,dt \qquad \cdots \quad (13.5.2)$$

and $\quad t' = t$

Hence we have

$$\frac{\delta}{\delta t} = \frac{\delta}{\delta t'} - u\frac{\delta}{\delta z}$$

and $$\frac{\delta}{\delta x} = \frac{\delta}{\delta z}$$

Using these transformations, equation (13.5.1) becomes

$$\mathrm{J} = qnv - q\mathrm{D}\frac{\delta n}{\delta z} - u\varepsilon\varepsilon_0\frac{\delta \mathrm{E}}{\delta z} + \varepsilon\varepsilon_0\frac{\delta \mathrm{E}}{\delta t'} \qquad \cdots \quad (13.5.3)$$

and Gauss' theorem in the new moving frame becomes

$$\varepsilon\varepsilon_0\frac{\delta \mathrm{E}}{\delta z} = q(n - \mathrm{N}) \qquad \cdots \quad (13.5.4)$$

Equation (13.5.3) can now be written using the simplification

$$\frac{\delta}{\delta z} = \frac{\delta}{\delta \mathrm{E}}\cdot\frac{\delta \mathrm{E}}{\delta z} = \frac{q(n-\mathrm{N})}{\varepsilon\varepsilon_0}\frac{\delta}{\delta \mathrm{E}}$$

as $$\frac{\mathrm{J} - \varepsilon\varepsilon_0\dfrac{\delta \mathrm{E}}{\delta t'} - q\mathrm{N}u}{qn} = (v(\mathrm{E}) - u) - \frac{\mathrm{D}q}{\varepsilon\varepsilon_0 n}\frac{\delta n}{\delta \mathrm{E}}(n - \mathrm{N}) \qquad \cdots \quad (13.5.5)$$

Assuming the diffusion constant to be field independent and integrating (13.5.5) from the external field E_0 to the highest field E_m gives the integral relation

$$\int_{\mathrm{E}_0}^{\mathrm{E}_m}\left(\frac{\mathrm{J} - \varepsilon\varepsilon_0\dfrac{\delta \mathrm{E}}{\delta t'} - q\mathrm{N}u}{qn}\right)\delta\mathrm{E} - \int_{\mathrm{E}_0}^{\mathrm{E}_m}(v(\mathrm{E}) - u)\delta\mathrm{E}$$

$$= -\frac{q\mathrm{N}\mathrm{D}}{\varepsilon\varepsilon_0}\int_{n(\mathrm{E}_0)}^{n(\mathrm{E}_m)}\delta\left(1 - \frac{\mathrm{N}}{n}\right)\frac{\delta n}{\mathrm{N}} \qquad \cdots \quad (13.5.6)$$

Consider the case of a stable domain *i.e.*,

$\frac{\delta E}{\delta t} = \frac{\delta E}{\delta t'} = 0$ when the moving frame is steadily moving with the stable domain. The maximum field E_m is the point where $\frac{\delta E}{\delta z} = 0$ *i.e.*, where the concentration is equal to the doping density and the charge changes sign from accumulation to depletion. The integration can be performed either from left to right or from right to left for E_0 to E_m.

In either case the integral on the R.H.S. vanishes. The second integral on the left is entirely a function of E and is independent of the path of integration while the first integral has a constant in the numerator and a variable in the denominator. This variable n is greater for the path through the accumulation while it is less than N for the path through the depletion. The only alternative left for this integral to be independent of path of integration is that $(J - qNu) = 0$. In this situation equation (13.5.6) gives

$$\int_{E_0}^{E_m} (\nu(E) - u)\, dE = 0 \qquad \cdots \quad (13.5.7)$$

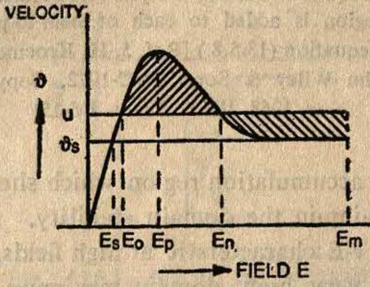


Fig. 13.5.1. Shows the fields and the velocities (when diffusion is ignored) within the matured domain, E_m is the maximum field in the domain, where space charge is zero. E_0 is the outside field on either side of the domain. E_n is the field at the highest charge concentration point of the domain. (All these assumes a matured domain moving with velocity U). $(E_d = E_o - E_s)$.

Fig. (13.5.1) shows the stable domain velocity u as related to the ν-E characteristic. Equation (13.5.7) implies that in this figure, the peak domain field E_m and the external field E_0 should be such that the areas of the cross hatched portions are equal. This requirement obviously defines a curve shown chain dotted in the figure (13.5.2), which gives the peak field E_m of a stable domain. The field independence of the diffusion constant in equation (13.5.5) has resulted in equal velocities for carriers at peak field position within the domain as for carriers outside the domain. The velocity (cf. fig. 13.5.2) V_n corresponding to the field E_n refers, therefore, to the highest concentration point of the accumulation region which should move with the same velocity to maintain the domain stability. Again considering the flatness of the v-E characterstic at high fields, $v_n = u$ is close to the velocity v_s for very high domain maximum fields. For such cases, we can find an approximate value for the maximum field E_m in terms of $(E_o - E_s)$ where $E_s = (v_s/\mu_o)$. By adding the double crossed region as shown in fig. (13.5.2) to both the equal areas and approximating the area under the v-E curve over the line indicated by v_s by the rectangular area we have

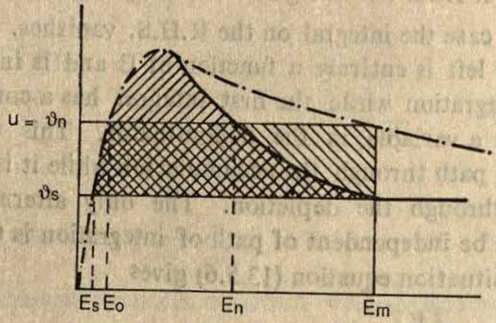


Fig. 13.5.2. Shows the equal area of the single hatched regions as required—Butcher's Equal Area Rule. Double hatched region is added to each of them to prove the identity of equation (13.5.8.) [Ref. 5, H. Kroemer, Copyright : John Wiley & Sons. INC-1972, Copyright © Nov. 1968, IEEE. Trans. Ed-15]

$$\int_{E_s}^{\infty} (v(E) - v_s)dE \approx (v_n - v)_s(E_m - E_s) \qquad \cdots \quad (13.5.8)$$

Defining

$$E_o^2 = \frac{1}{\mu_o}\int_{E_s}^{\infty} (v(E) - v_s)\, dE \qquad \cdots \quad (13.5.9)$$

and expressing v_n and v_s in terms of the low field mobility μ_0, we have for eqn. (13.5.8)

$$E_c^2 = (E_m - E_s)(E_0 - E_s) \qquad \cdots \quad (13.5.10)$$

wher E_c is a quantity entirely dependent on the materials $v(E)$ characteristic alone and not on the domain details. For GaAs Ruch and Kino's results give $E_c = 2{\cdot}935$ KV/cm.

13.6 Domain Mode

The important consequence of the discussion of the equal area rule points out that the drift velocity at the highest field is always slightly less than the drift velocity outside the domain. Since the electron density at these locations are same also, the currents have to be same while diffusion is present. Hence diffusion current should make up for this velocity difference. Hence the domains accumulation and depletion regions must be adjacent to each other. There cannot exist a neutral high field extended region between the two. Hence the domain field must be about triangular not trapezoidal as shown in fig. (13.6.1).

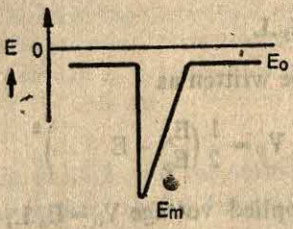


Fig. 13.6.1. Shows the triangular shape of the domain field.

The field rise on the accumulation region must be much steeper than the field decrease on the depletion region. This is because the space charge density in the depletion side is limited to N, the doping level, while in the accumulation region charge density can rise to any value. Hence we have here a situation like a metal semiconductor barrier region where all of the voltage drops across the depletion region. Assuming complete depletion of the domain, we can write for the domain maximum field E_m

$$E_m = \left(\frac{2qNV_d}{\varepsilon\varepsilon_0}\right)^{\frac{1}{2}} + E_0 \qquad \cdots \quad (13.6.1)$$

In terms of the applied voltage, the domain voltage

$$V_d = V_a - E_0 L = (E_b - E_0) L \quad \cdots \quad (13.6.2)$$

where E_b is the average bias field and $E_b L = V_a$.

From (13.6.2) with (13.6.1)

$$E_b = E_0 + \frac{1}{2E_h}(E_m - E_0)^2 \quad \cdots \quad (13.6.3)$$

where E_h is the highest field that can be supported by the ionised donors of the entire device length and is given by

$$E_h = (qNL/\varepsilon\varepsilon_0) = 1 \cdot 45 \times 10^5 \frac{V}{cm} \times (NL/10^{12} \text{ cm}^{-2}) \quad \ldots \quad (13.6.4)$$

note that it is proportional to the NL product of the device. The two fields E_m end E_0 are related by the equal area rule.

Equation (13.6.1) gives for the domain voltage

$$V_d = (\varepsilon_0 \varepsilon L/2qNL)(E_m - E_0)^2 - (L/2E_h)(E_m - E_0)^2 \quad \cdots \quad (13.6.5)$$

using $\omega_c - (qN\mu_0/\varepsilon\varepsilon_0) \quad \cdots \quad (13.6.6)$

and $V_p = E_p . L$

Equation (13.6.5) can be written as

$$(V_d \, \omega_c L/v_p \, V_p - \frac{1}{2}\left(\frac{E_m}{E_p} - E \quad \right)^2 \quad \cdots \quad (13.6.7)$$

For a certain applied voltage $V_a = E_b . L$, a voltage $E_0 . L$ drops in the sample due to low field E_0, while the rest remains as excess voltage V_d across the domain. From equation (13.6.7), when $E_0 = E_p$, equal area rule demands (cf. fig. 13.5.1) $E_m = E_p$; hence V_d, the excess domain voltage is then zero. Equation (13.6.2) with equation (13.6.7) indicates that as the bias voltage increases the excess domain voltage also increases and this can go on till domain field causes semiconductor breakdown in the domain region. That is the domain voltage can increase but that the value is decided by the bias voltage, peak velocity v_p of the velocity field characteristics and $\omega_c L$ *i.e.*, the NL product of the device. For a certain NL product the variation of the domain voltage versus the field external to domain E_0 can be plotted from equation (13.6.7) as shown in fig. (13.6.2). So long as the domain is present the anode current is decided by E_0 and in view of equation (13.6.2) and fig. (13.6.2), corresponding to a stable domain voltage the $(E_0.L)$ voltage drop multiplied by the scale factor $K=(\omega_c L/v_p V_p)$, can be added to obtain the point corresponding to the applied voltage. This locates the

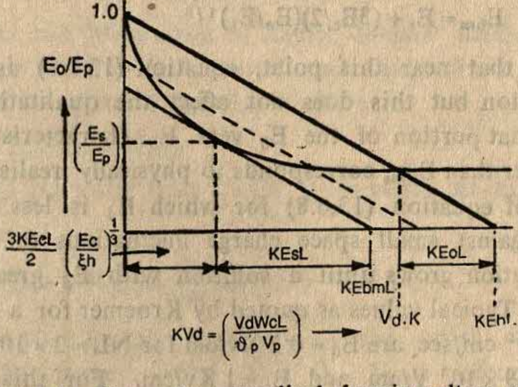

1.0
E_0/E_p
$\left(\frac{E_s}{E_p}\right)$
$\frac{3KE_cL}{2}\left(\frac{E_c}{\xi h}\right)^{\frac{1}{3}}$
KEsL
KEoL
KEbmL
Vd.K
KEhl.
$KVd=\left(\frac{VdWcL}{\vartheta_p V_p}\right)$

Fig. 13.6.2. Shows the normalised domain voltage as a function of normalised field outside the domain. The load line is drawn using the seale factor K corresponding to the NL product. The curve also shows the dependance of E_s on E_{hr} hence the NL product. Note that after domain nucleation as E_b decreases E_o increases since KE_oL increases in size.

load line as shown in fig. (13.6.2). If the stable domain exists while the applied voltage is decreased, the load line can be shifted parallel to itself. The excess domain voltage decreases while E_0L voltage drop increases (cf. fig. 13.6.2 dotted lines) indicating that current increases. Ultimately, the stable domain can no longer exist if the applied voltage falls below a certain value.

In order to find this minimum bias voltage or the corresponding bias field E_{bm}, write equation (13.6.3) using equation (13.5.10) as

$$E_b=E_s+E_d+[(E_c^2/E_d)-E_d]^2/2E_h \quad \cdots \quad (13.6.8)$$

where $E_d=(E_0-E_s)$.

From fig. (13.5.2), it is seen that E_d is very small compared to all other terms in this equation; so, equation (13.6.8) can be approximated to

$$E_b=E_s+E_d+(E_c^4/2E_h.E_d^2) \quad \cdots \quad (13.6.9)$$

Equation (13.6.9) shows that the bias field has a minimum value E_{bm} above which a mature domain can exist but none below. This occurs for value of

$$E_{dm} = E_c(E_c/E_h)^{1/3} \qquad \cdots \quad (13.6.10)$$

when

$$E_{bm} = E_s + (3E_c/2)(E_c/E_h)^{1/3} \qquad \cdots \quad (13.6.11)$$

Note that near this point, equation (13.5.8) is not a good approximation but this does not affect the qualitative behaviour. Actually, that portion of the E_b vers. E_0 characteristic for which E_d is greater than E_{dm}, corresponds to physically realisable solution. Solutions of equation (13.6.8) for which E_d is less than E_{dm} are unstable against small space charge fluctuations (cf. ref. 5). Any such fluctuation grows until a solution with E_d greater than E_{dm} develops. Typical values as quoted by Kroemer for a drift velocity of $1{\cdot}16 \times 10^7$ cm/sec, are $E_d = 0{\cdot}2$ Kv/cm for $NL = 2 \times 10^{12}$ while one has $E_h = 2{\cdot}9 \times 10^5$ V/cm and $E_s = 1$ Kv/cm. For this NL product from the approximate expressions, we have $E_{dm} = 635$ V/cm and $E_{bm} \sim 2$ Kv/cm, *i.e.*, approximately E_0 is within a factor of two from E_s. The above estimate of Kroemer suggests $E_0 \sim E_s$ of fig. (13.5.2).

From the above discussion, we see that once a stable domain has formed there exists a minimum domain sustaining voltage $E_{dm}.L$. If the bias voltage drops below $E_{bm}.L$, the domain cannot exist. The domain nucleation does not occur until the bias voltage is greater than $E_v.L$. Once the domain is nucleated the bias field

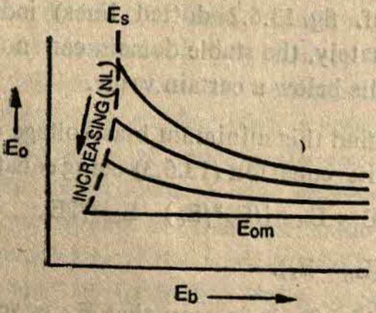


Fig. 13.6.3. Shows the bias field *vrs.* characteristics for different NL products. Note that corresponding to E_{dm} for each NL product, an E_{om} is expected.

can be reduced to the value given by equation (13.6.11), obviously a much lower value than required for the peak velocity field E_p. The bias field E_b versus E_0 can be plotted from equation (13.6.9) and is shown in fig. (13.6.3). From equation (13.6.11) it is evident that E_{bm} decreases with increase of NL.

Domain can breakdown at higher domain fields. Hence this sets an upper limit to the highest domain field. From eqn. (13.6.3), neglecting E_0 and putting $E_m = E_{br}$, we have

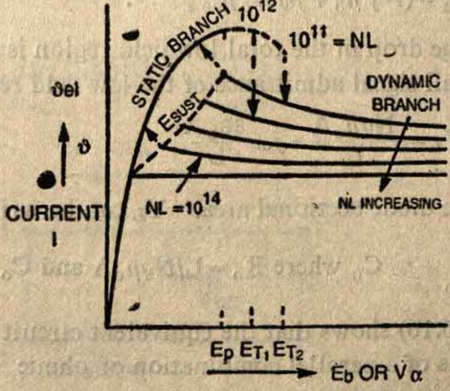
$$E_b = (E_{br}^2/2E_h) \quad \cdots \quad (13.6.12)$$

In GaAs, E_{br} is greater than 2×10^5 V/cm hence the highest average bias field

$$E_b = 1{\cdot}38 \times 10^5 \frac{V}{cm}(10^{12}\ cm^{-2}/NL) \quad \cdots \quad (13.6.13)$$

Thus we see that for NL products of the order of critical NL product, the highest bias field is large compared to $E_p = 3{\cdot}2 \times 10^3$ V/cm.

The knowledge of the steady state domain parameters allows one to determine the time dependent current-voltage or velocity-average bias field characteristics. The characteristic will obviously consist of a static and a dynamic branch, as shown in fig. (13.6.4) from computer simulation results (cf. ref. 10). The static part corresponds to the quasi-ohmic range of the velocity field characteristics. This continues upto the threshold field E_T which is little



Fig, 13.6.4. Shows the dynamic paths of domain formation and extinction greater than E_p and depends on the NL product, being greater for smaller NL product. At E_T, a domain is formed when the drift current drops to the dynamic branch. This state continues as long as the domain is in transit. If the bias field falls below the sustaining field E_{bm} while the domain is propagating, the domain is quenched and the drift velocity reverts to the static branch, as shown by the arrow in fig. (13.6.4).

13.6.1 Small signal impedance

The small signal impedance of the diode can be found out as follows. For the low field ohmic region, the current can be written as

$$J = Nq\,\mu_0\,E + \frac{\varepsilon\varepsilon_0\,dE}{dt} \qquad \cdots \quad (13.6.14)$$

Assuming that the domain occupies a fixed spatial distance d and the total current is composed of a time varying a.c. part superposed on a d.c. current, while the field outside the domain similarly has time varying a.c. fluctuation superimposed on the low field regions d.c. part E_0 we have for small signals $J = J_0 + J_1(t)$ while $E = E_0 + E_{01}(t)$. This time varying part of the field is a sinusoidal variation of the form $E_{01}(t) = E_{01}e^{j\omega t}$. The time-varying part of the current for low field region is of the from $J_1 \exp j\omega t$ hence substituting for J and E in equation (13.6.14) and separating the d.c. and a.c. parts, we have the small signal current as

$$J_1 = (Nq\,\mu_0 + j\omega\varepsilon\varepsilon_0)E_{01} \qquad \cdots \quad (13.6.15)$$

The a.c. voltage drop in the total low field region is given by $v = LE_{01}$. Hence the small signal admittance of the low field region is

$$Y_L = \frac{Nq\mu_0 A}{L} + j\omega\frac{\varepsilon\varepsilon_0 A}{L} \qquad \cdots \quad (13.6.16)$$

where A is the diode sectional area. Y_L can be written as

$$Y_L = \frac{1}{R_0} + j\omega\,C_0 \text{ where } R_0 = L/Nq\mu_0 A \text{ and } C_0 = \varepsilon_0\varepsilon A/L$$

Equation (13.6.16) shows that the equivalent circuit of the low field region consists of a parallel combination of ohmic resistance and a capacitance.

On the other hand, if the diode terminal voltage varies when the domain is present, the current changes so the total a.c. resistance can be obtained from

$$\frac{L\,dE_b}{A\,dJ} = \frac{L}{A}\,\frac{dE_b}{dE_d}\cdot\frac{dE_d}{dJ}$$

which can be written using equation (13.6.9) as

$$\frac{dV_a}{dI} = \frac{L}{A}\,\frac{dE_0}{dJ}\left(1-\frac{E_c^4}{E_h E_d^3}\right)$$

while E_0 being in the low field region, we have

$$\frac{dV_a}{dI} = \frac{L}{\mu_0 q NA}\,(1-E_c^4/E_h E_d^3) \qquad \cdots \quad (13.6.17)$$

Equation (13.6.17) shows that when the domain is charging or discharging the rest of the region has a positive resistance $(L/\mu_0 qNA)$ while the domain has a negative resistance given by

$$R_d = -\frac{L}{\mu_0 q\, NA}\,\frac{E_c^4}{E_h\, E_d^3} \qquad \cdots \quad (13.6.18)$$

From (13.6.1) the domain charge variation can be represented as due to domain voltage variation as

$$\delta\, V_d(t) = \frac{\varepsilon\varepsilon_0}{qN}\,(E_m - E_0)\;\delta(E_m - E_0)$$

the fields E_m as also E_0 varies during domain voltage variation. If $E(t)$ is substituted for time varying $(E_m - E_0)$, we have

$$\delta\, V_d(t) = \frac{d(t)}{\varepsilon\varepsilon_0}\,[\varepsilon_0\varepsilon\;\delta\, E(t)]$$

where $d(t)$ is the time dependent width of the triangular domain. Hence we have

$$\delta V_d(t) = \frac{d(t)}{\varepsilon\varepsilon_0}\,\frac{\delta Q(b) Q(t)}{A} = \frac{\delta Q(t)}{C_d(t)} \qquad \cdots \quad (13.6.19)$$

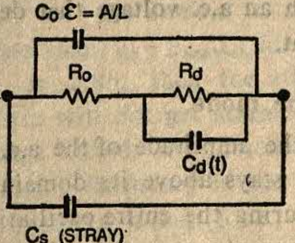


Fig. 13.6.5. Shows domain mode equivalent circuit for the Gunn diode.

where $\delta Q(t)$ and $C_d(t)$ are the time dependent charge and domain capacitance. The domain, therefore, represents a negative resistance in parallel with a capacitance. Hence the equivalent circuit for the diode can be drawn for the domain mode as shown in fig. (13.6.5)

It may be noted that under dynamic conditions of domain growth Kurokawa (cf. ref. 11) has shown that

$$\frac{\delta V_d}{\delta t} = \int_{E_0}^{E_m} [v(E_0) - v(E)]\, dE \qquad \cdots \quad (13.6.20)$$

which can give the equivalent circuit of the domain in more rigorous terms.

13.7 Resonant Circuit Loading

So far we have discussed the space charge instabilities under fixed bias voltage conditions. The Gunn diode would actually work in the MW range where the load would obviously be a MW resonant cavity or the like as shown in fig. (13.7). Hence we need to consider the voltage at the diode terminals as a d.c. bias superposed on which an a.c. voltage that develops across the resonant circuit is present.

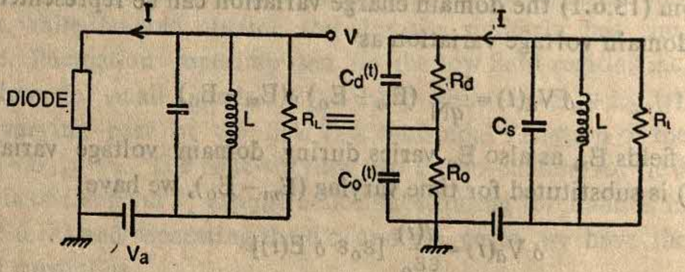


Fig. 13.7. Shows the diode equivalent circuit in parallel with the loaded resonant circuit. On the left the diode capacitances are lumped across the resonant circuit.

13.7.1 Transit time mode

As long as the amplitude of the a.c. voltage remains small such that the device stays above its domain nucleation threshold voltage $V_T = E_T.L$, during the entire oscillation cycle nothing fundamentally new happens. As soon as a domain is formed the anode current falls, remains nearly constant as long as the domain is in transit. As the domain is delivered to the anode, the current rises since the anode field rises, which immediately falls only to cater for the new domain's growth, the total diode bias field being above the threshold value V_T always. Hence the current pulses that are delivered after each domain transit time are as shown in fig. (13.7.1) These current pulses excite the resonant circuit and cause the voltage variation as shown in this figure. Because the domain moves faster when the domain voltage is low and slower when this voltage is high, a variation in the transit time is possible by varying the bias voltage and/or the amplitude of voltage swing. Thus some amount of tuning is possible in this mode which is known as the 'transit time mode'.

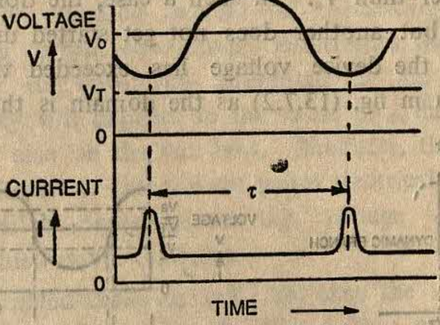


Fig. 13.7.1. Shows the transit time mode current 'I' and voltage 'V' referred to in fig. 13.7.1 above *vrs.* time.

13.7.2 Delayed mode

If the a.c. amplitude brings the total device voltage periodically below the domain nucleation threshold value V_T, either for a time long enough for the domain to reach the anode or to a value low enough (less than $V_s = E_{bm}.L$) for the space charge to collapse (quenched) or both, then the situation is different. In either case new domain will not get started and will be delayed until the device voltage exceeds the threshold voltage again. This means that the oscillation frequency is not determined by the transit times but by the resonant circuit load. In this way, operation at frequencies drastically different from the transit time frequency is obtainable. At frequencies lower than the transit time frequency, the delay provides the possibility of wide frequency tunability. For example, in the delayed mode, we assume that a domain is formed at $t=0$ and while it is in transit the r.f. voltage has made the device voltage fall below V_T but the device voltage remains greater than V_s. In such a case, the domain is delivered to the anode, but another does not get started until the time $t=\tau+\delta=T$ when the device voltage has exceeded the threshold V_T. This is shown in fig. (13.7.2) as the domain is thus delayed by a

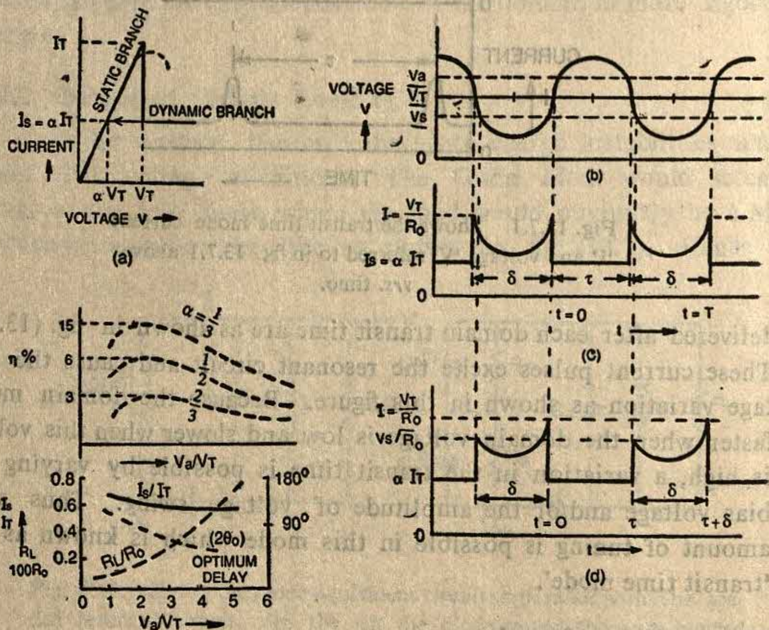


Fig. 13.7.2. Shows the ideal delayed domain mode. (a) State and dynamic parts of current voltage characteristics. (b) Shows the R.F. voltage relative to the voltage applied V_a, threshold V_T and sustaining V_s (c) Shows the current pulse shape for optimised generated power and the corresponding delay δ. (d) Current pulse shape, when δ is not optimised. (e) Shows the efficiency *vrs.* applied voltage normalised to $E_p.L=V_T$, I_s/I_T valley to peak current ratio, (R_L/R_o) load resistance normalised to low field device resistance and optimum delay *vrs.* normalised applied D.C. voltage.

time δ. it is assumed here that the domain collapses instantaneously on reaching the anode or matures instantaneously when the device voltage rises above V_T and the static current voltage characteristic is ideal, as shown in fig. (13.7.2a) in place of one of fig. (13.6.4). In this it is assumed that $V_p = V_T$.

For maximum output, maximum current pulse duration and height possible should be delivered to the resonant circuit as in class C oscillators. The characteristics of such oscillators are well-known where the standing current $\alpha.I_t$ drastically reduces the efficiency and the current wave shape determines the power. It is obvious that for the waveforms presented in fig. (13.7.2b), the optimum efficiency will depend on the value of α, the valley to peak current ratio, as also on the d.c. bias. Similarly, the d.c. current, optimum delay (equal to the current pulse electrical angle) and the optimum load will be dependent on d.c. voltage applied. These curves are presented in fig. (13.7.2c).

It is to be noted from fig. (13.7.2b) that the domain transit angle will be controllable by the bias voltage and will increase slightly with increase of this voltage. After the domain is delivered to the anode so long as the device voltage is below the threshold, a new domain is delayed in forming. The circuit frequency is decided by the transit time T plus the delay δ. For a d.c. bias equal to V_T circuit tuning can accommodate a circuit frequency such that the δ can vary between zero to half the circuit period. In other words, the circuit frequency can be varied by as much as factor of two between the transit time frequency $f_T = \frac{1}{\tau}$ when $\delta = 0$ to $0{\cdot}5 f_T$, when $\delta = \tau$. Obviously, the output power, efficiency even optimum load will vary with it. For d.c. bias higher than V_T, the circuit tuning gets narrower than 2 : 1. It is to be noted that, in this mode some correlation of circuit frequency to the sample length is required. It is possible to sustain oscillations for bias voltages less than the threshold voltage provided the a.c. voltage brings the diode voltage above threshold and the oscillations generate enough power to be self-sustaining. If the oscillations are not self-sustaining, by triggering with a suitable transient, the diode can be used as a monostable.

13.7.3 Quenched mode

At frequencies higher than the transit time frequency the total diode voltage can become too low before the domain has been delivered to the anode. In such cases, if the diode voltage is lower than the sustaining voltage $E_sL = V_s$, then the domain starts to discharge or quench. We assume here that quenching occurs instantaneously and draw the ideal current voltage waveforms for this mode as shown in fig. (13.7.3). After the domain is quenched, the diode behaves as an ohmic resistance until the diode voltage has exceeded the V_T again when a new domain nucleates and starts travelling as a mature domain. The growth period of the domain to maturity has been ignored in drawing this figure. It is seen again the valley to peak current ratio α affects the optimum efficiency. It would be seen from fig. (13.7.3) that in general the efficiency is more than a factor of two

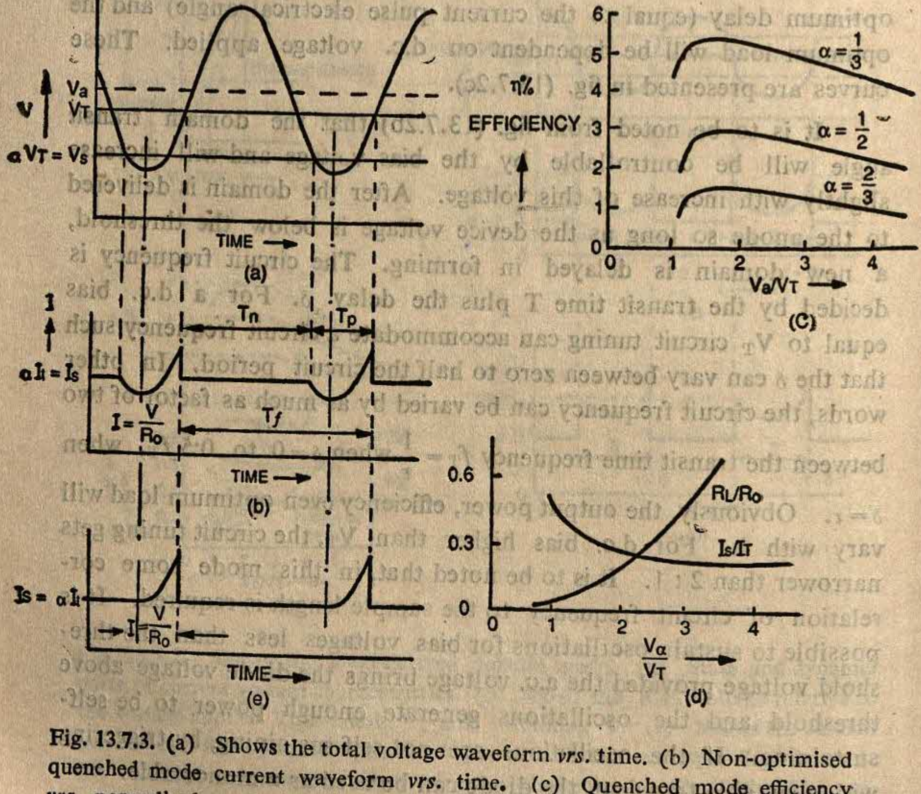


Fig. 13.7.3. (a) Shows the total voltage waveform *vrs*. time. (b) Non-optimised quenched mode current waveform *vrs*. time. (c) Quenched mode efficiency *vrs*. normalised applied voltage. (d) Normalised load and valley to threshold current ratio *vrs*. applied voltage. (c) Current waveform when domain is quenched just at the nagative peak of the R.F. voltage.

lower than the delayed mode. This figure is based on ref. (cf. 12). The corresponding optimum load is higher by about the same factor than in the delayed mode. It would be seen from the figure that for quenched mode operation the circuit period T_f is less than transit time (is required). So, Gunn effect oscillations at frequencies much higher than the transit time frequency is possible. But it is to be remembered that neither quenching nor the growth are instantaneous. Since it depends on NL product it has to be high compared to non-quenched operation.

Quenched mode operation as seen from figure (13.7.3a) is rather inefficient. The portion of the device that lies beyond the quenching distance acts as a purely positive series resistance throughout the r.f. cycle. Obviously, this portion of the device could be eliminated so that the transit time frequency would coincide with the desired frequency of the quenched mode. Further comparison of figures (13.7.3) with (13.7.2) will show that the quenched mode power output is also much lower than in the delayed domain mode. Therefore, the main interest in quenched opreation is for application like signal generators, where wide band tuning is required rather than power.

13.7.4 LSA mode

Next we discuss the case of quenched accumulation layer mode. It would be seen that quenching of AL needs less field lowering from the threshold than in the domain mode. AL quenching time can also be shorter than the domain quenching time. Finally, in contrast to the quenched domain mode the portion of the device between the point of quenching to anode remains in the negative mobility region for most of the oscillation cycle, except for only a brief period, between quenching and reinjection of an AL. Such a negative resistance generates oscillation power rather than absorbing as in a positive resistance for this region, in domain mode. Hence, it is expected that quenched AL mode would give higher power at high frequencies.

It may be noted that in quenched AL mode, during the early part of the travel of AL towards the anode and when the AL is near the anode, the magnitude of the field ($E_b + E^+$) on the anode side of the AL is not very much different from the bias field. Hence the total bias field has to drop only slightly below the threshold to bring the device in the positive mobility region thus enabling faster AL quenching during these periods. The field remaining near the threshold keeps the current high during quenching. As a result, the quenched AL mode has a much higher efficiency than the quenched domain mode, if the oscillation frequency can be made high compared to the transit time frequency. If this frequency is very high compared to the transit time frequency the AL proceeds only a very short distance, where the AL starts to quench since the voltage has become low. Before it is completely quenched another AL starts since the R.F. has changed phase by this time. This AL and the residue of the earlier one starts growing, then a short time later, they both start quenching. Thus a series of gradually decaying ALs can be envisaged. In this case, most of the device remains in the negative resistance region. For this situation, the mode is named Limited Space-charge Accumulation (LSA) mode. Ideally one can assume uniform field over most of the diode length.

The LSA mode of operation can be analysed analytically by assuming a three-piece linear approximation for the velocity field characteristic as shown in figure (13.7.4) and the structural requirements can be found out (cf. ref. 5, 6, 12). In this figure, it has been assumed that the sinusoidal R.F. field ($E_p.b_1$) is superposed on the bias field $E_p b_0$. As seen from the figure, the growth of space-charge would occur over an electrical angle $2\theta_n$ while decay will occur over an angle $2\theta_p$ and the negative differential mobility can be taken as a constant. Since in AL mode the growth of space-charge is accompanied with a rise of the anode field while the space-charge decay correspondingly causes this field to decrease, the current wave shape can be figured from the corresponding J-E curve obtained from the carrier velocity at the anode using the expression $J = qNv$. Fourier analysis of this current waveform gives the d.c. and the fundamental current components. The d.c. current with the bias voltage gives the power input while the fundamental current with the R.F. voltage gives the R.F. power output. So, the efficiency can be calculated for this mode. This line of reasoning is left to the reader as an exercise (cf. Problem 13.7). Here we do not follow piece-wise linear approximation since the negative differential mobility varies over wide range; so one should speak of a time average of this mobility. This would obviously depend on the time spent by the signal in these regions (cf. problem 13.9); hence, on the waveform too. We need to remember that during the finite quenching time, space-charge cannot be completely quenched, only that it would be attenuated heavily leaving a small residue. In the next cycle, when the voltage exceeds the threshold again, this residue will grow while a new AL will get introduced at the cathode. Hence,

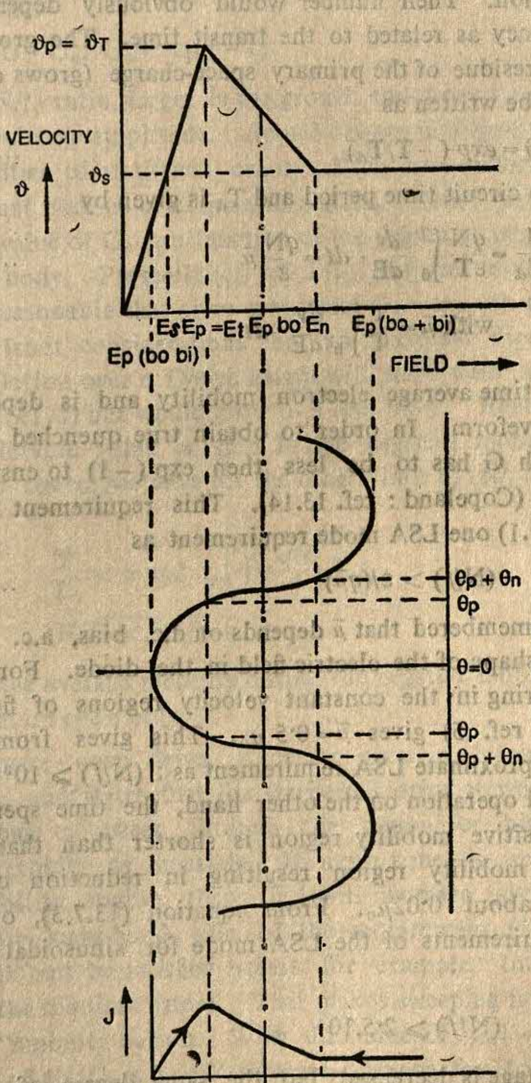


Fig. 13.7.4. LSA mode idealised sinusoidal R.F. instantaneous field velocity diagram for $E = E_p\,(b_0 - b_1 \cos \omega t)$

for true LSA operation the regrowing residue must remain weak compared to the newer ones. There will be series of gradually diminishing space-charge accumulation layers moving along the diode body and the whole diode body would be in the negative resistance region. Their number would obviously depend on the circuit frequency as related to the transit time. The growth factor by which the residue of the primary space-charge (grows or) decays per cycle can be written as

$$G = exp\,(-T/T_a) \qquad \cdots \quad (13.7.1)$$

where T is the circuit time period and T_a is given by

$$\frac{1}{T_a} = \frac{qN}{\varepsilon T}\int_0^T \frac{dv}{dE}\cdot dt = \frac{qN}{\varepsilon}\,\bar{\mu} \qquad \cdots \quad (13.7.2)$$

$$\text{with } \bar{\mu} = \frac{1}{T}\int_0^T \frac{dv}{dE}\cdot dt$$

where $\bar{\mu}$ is a time average electron mobility and is dependent on the signal waveform. In order to obtain true quenched operation, overall growth G has to be less then $\exp(-1)$ to ensure space-charge decay (Copeland : ref. 13.14). This requirement gives with equation (13.7.1) one LSA mode requirement as

$$(N/f) > \varepsilon/(q\bar{\mu}) \qquad \cdots \quad (13.7.3)$$

It is to be remembered that $\bar{\mu}$ depends on d.c. bias, a.c. amplitude and the wave shape of the electric field in the diode. For a square wave not entering in· the constant velocity regions of fig. (13.7.4), Kroemer (cf. ref. 5) gives $\bar{\mu} = 0{\cdot}5\,\mu_0$. This gives from equation (13.7.3), an approximate LSA requirement as : $(N/f) \geqslant 10^8$ for GaAs. For sinusoidal operation on the other hand, the time spent by the R.F. in the positive mobility region is shorter than that spent in the negative mobility region resulting in reduction of average mobility to about $0{\cdot}02\mu_0$. From equation (13.7.3), one of the structural requirements of the LSA mode for sinusoidal operation becomes,

$$(N/f) \geqslant 2{\cdot}5.10^4 \qquad \cdots \quad (13.7.4)$$

If the bias voltage is decreased, but the same degree of quenching is assumed, the R.F. voltage spends more time in the positive mobility region thus increasing $\bar{\mu}$. By equation (13.7.3), even a lower N/f can then provide the required degree of quenching, In addition to the lower limit of (N/f) mentioned above, there is an upper limit for the (N/f) ratio for LSA operation. This comes from the fact that the growth G_n during the negative mobility region must be related to the decay G_p during the positive mobility region such that

$$G = G_n . G_p = \exp(-1) \qquad \cdots \quad (13.7.5)$$

Larger the (N/f) ratio, larger is the growth and decay type oscillation of the space-charge amplitude. Actual devices will have unavoidable non-uniformities of statistical origin. Such non-uniformities may act as incipient nucleus of domian formation. Hence the maximum permissible value of G_n is a function of the degree of non-uniformity in the diode body. Presently (cf. ref. 14), 1 per cent non-uniformity is taken as reasonable that gives a value of $G_n \sim \exp(5)$ as tolerable. This arises from considerations that an incipient dipole domain with accumulation over a Debye length will start with a voltage of the order of (kT/q) but would grow to a few volts. Hence taking a maximum tolerable value of gain in the negative mobility region, as mentioned above, we have for the other limit of (N/f) ratio from equations

$$\frac{T}{T_n} < 5 \text{ and } \left(\frac{1}{T_n}\right) = \frac{qN}{\varepsilon}\left|\bar{\mu}_n\right|$$

as $$(N/f) < 5\varepsilon/(q|\bar{\mu}_n|) \qquad \cdots \quad (13.7.6)$$

where $\bar{\mu}_n$ is the average mobility during the negative mobility portion of the oscillation cycle only.

For square wave drive, Kroemer gives a lower limit for $|\bar{\mu}_n| = 0{\cdot}01\ \mu_0$ and states that for sinusoidal drive it would require a high d.c. bias to sweep through the region of high negative mobility as rapidly as possible. If more time is spent in the negative mobility region, then $|\bar{\mu}_n|$ will increase, giving a lower value for the right-hand side of the requirement (13.7.6). The condition will not be satisfied unless, for example, the frequency increases to the required extent. That makes sweeping faster through the negative mobility region. With this value of $|\bar{\mu}_n| = 0{\cdot}01\ \mu_0$, the upper limit of (N/$f$) is 2×10^5 sec/cm^3. Hence for LSA mode we have the structural limits as,

$$2{\cdot}5 \times 10^4 \frac{\text{sec}}{\text{cm}^3} < (N/f) < 2 \times 10^5 \frac{\text{sec}}{\text{cm}^3} \qquad \cdots \quad (13.7.7)$$

If we have a device with $NL = 2 \times 10^{12}$ that is working in the transit time mode, the transit time is $1/f = L/10^7$ seconds. For this device $N/f = 2 \times 10^5 \frac{\text{sec}}{\text{cm}^3}$. Hence at least in the lower portion of the (N/*f*) range, the LSA mode devices are less heavily doped than transit time devices.

Define space-charge overall decay $G = \exp(-\delta)$ where $\delta = 1$ was used in the above discussions. Carroll (cf. ref. 12) gives analytically determined normalised bias field versus efficiency curves from his piece-wise linear approximation of v-E curves for sinusoidal R.F. voltage as shown in fig. (13.7.5) for $\delta = 1$. This author gives the (N/*f*) ratio and efficiency versus normalised bias field from analytical methods, as shown in fig. (13.7.6), for zero value of δ as well. Note

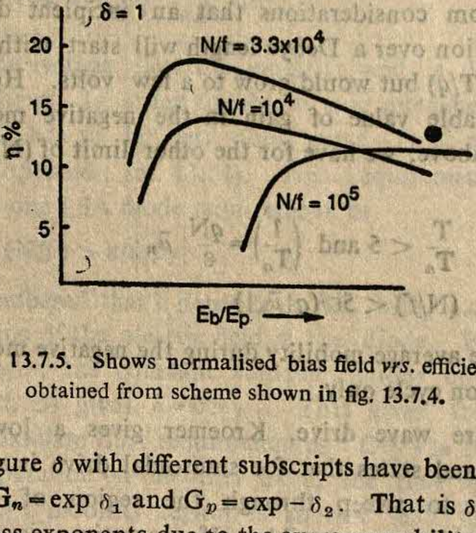


Fig. 13.7.5. Shows normalised bias field *vrs.* efficiency obtained from scheme shown in fig. 13.7.4.

that in this figure δ with different subscripts have been used. These are defined as $G_n = \exp \delta_1$ and $G_p = \exp - \delta_2$. That is δ_1 and δ_2 are the gain and loss exponents due to the average mobility values during only the negative and only the positive mobility portions of the oscillation cycle respectively. We have for one complete cycle, $G = \exp(\delta_1 - \delta_2)$. Zero value for δ can be achieved with $\delta_1 = \delta_2$. We have seen that the highest permissible value of δ_1 is 5 while the lowest value of $\delta = 1$ is not regarded to give useful growth (cf. art. 13.2). Fig. (13.7.6) shows that the useful N/*f* range increases with increase of the bias field. From these curves, as also from our previous discussions, it can be concluded that the LSA mode, in general, requires high bias field. Typical load impedance lies in the range of 10 to

100 times the low field resistance of the device. For optimised bias field and load resistance, the efficiency of LSA mode can reach as high as 17 per cent for bias fields between 10—15 KV/cm in GaAs.

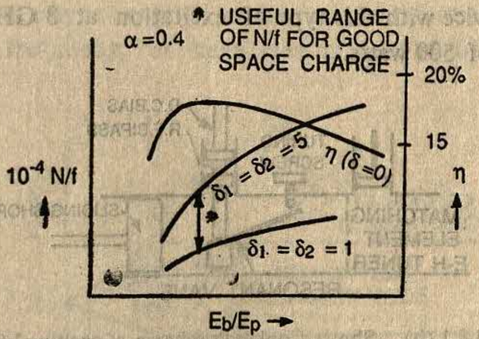


Fig. 13.7.6. Shows normalised bias field *vrs.* N/*f* ratio and efficiency.

13.8 Starting of LSA Mode

We have seen that the LSA mode requires high bias field of operation. Unless the R.F. swing develops to large amplitude, it is not possible to start the LSA. When the AL mode starts to develop to large amplitude, small doping non-homogeneity may change it to domain mode of operation. If domain is formed, then it will absorb large voltage in its formation. If the applied voltage is large, domain peak field may exceed, E_{br} when the diode will break down. Hence the first requirement is good quality of crystal with possibly less than one per cent doping non-uniformity. With such good quality of crystal, one needs to build up the R.F. voltage very rapidly to large amplitude to obtain the LSA mode of operation.

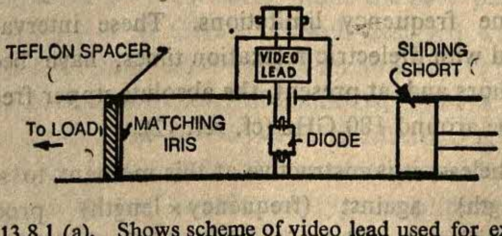


Fig. 13.8.1 (a). Shows scheme of video lead used for exciting LSA mode. [W.K. Kennedy, L.F. Eastman & R.J. Gilbert. © Copyright 1967 IEEE, Trans. ED-17]

One approach is to use very fast rising pulses to the sample through a suitably designed video lead as shown in fig. (13.8.1a). That is to impart R.F. energy into the cavity that causes the cavity to ring, thus helping the R.F. voltage to grow. Kennedy (cf. ref. 15) has reported a device with this type of excitation at 8 GHz giving short pulse power of 500 watts.

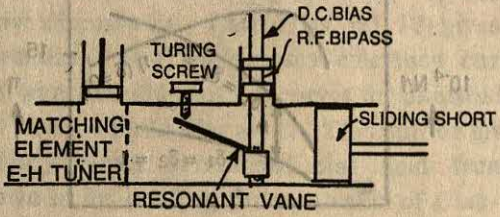


Fig. 13.8.1 (b). Shows Copland's scheme of exciting LSA mode.

A resonant vane formed a lightly loaded high Q resonant circuit across the diode (cf. fig. 13.8.1b) in Copeland's work (ref. 16). The initial transients are produced by the domains. The load matching elements are kept far away from the resonant cavity. So, by the time the diode perceives its presence by reflections from the matching element, the R.F. voltage builds up to large value because the resonant cavity is, as if, not loaded. It has been seen by Heeks (ref. 17) that when high frequency LSA oscillations build up, low frequency oscillations below the waveguide cut-off exist below the resonant vane which forms a capacitance with the guide, the guide acting as inductance below cut-off. This low frequency component helps in the high frequency formation above the wave guide cut-off frequency.

13.9 Mode Chart of Gunn Devices

In multivalley semiconductors, the intervalley scattering rates provide some frequency limitations. These intervalley transition rates, coupled with dielectric relaxation times, have been considered by many authors and, at present, the absolute upper frequency limit is assumed to be around 180 GHz (cf. ref. 12).

Nevertheless, it is instructive at this moment to summarise the (doping × length) against (frequency × length) product requirement for the different modes of working of Gunn diodes discussed so far. This is shown in fig. (13.9.1). Self-sustained oscillations require

NL products greater than 10^{12}. For NL products between 10^{11}-10^{12}, some amount of per cycle gain may be expected when the Gunn diode is used as small signal amplifier, particularly around the transit time frequency. For NL products greater than 10^{12}, some amount of frequency tuning is expected in the transit time mode. Below the transit time frequency, we have seen in delayed domain mode, frequency tuning is possible down to about half of the transit time frequency. The quenched domain mode of operation occurs above the transit time frequency range. Finally, the LSA mode of operation defined by (N/f) condition (13.7.7), is shown between two parallel lines in fig. (13.9.1).

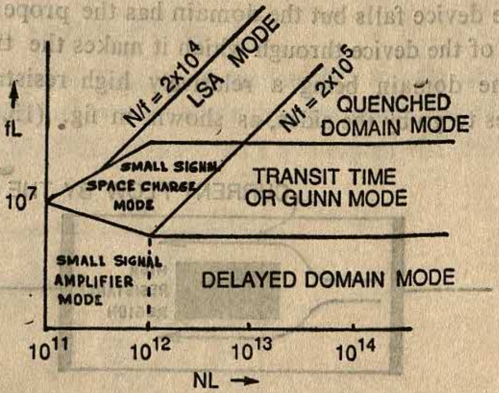


Fig. 13.9.1. Shows the mode chart of Gunn diodes. [Ref. 3. John A. Copland. Copyright © Sept. 1967 IEEE, Trans. ED-14]

Best CW power frequency squared values obtained with Gunn devices are around $5{\cdot}3 \times 10^2$ W GHz^2 while the best pulsed value for this product is $9{\cdot}8 \times 10^4$ W GHz^2. In well designed waveguide cavity, the temperature variation can be less than 50 KHz/°C between -40°C to $+70$°C. The AM noise level of Gunn Oscillator is quite low while FM noise is comparable to klystron oscillators. Amplifier noise figure NF of the order of 30 *d*B has been quoted by some authors. Gunn devices can be used as reflection type amplifiers also in switching and waveform generating functions. This is the topic of our discussion in the next section.

13.10 TED Signal Processing

Around 1967 onwards it was realised that dipole domain formation associated with current changes in the external circuit could be used for subnanosecond pulse generation, waveform generation and signal processing (cf. ref. 18). One principle of pulse and waveform generation lies in the fact that when a domain has formed, the current through the device falls but the domain has the property of filling the whole area of the device through which it makes the transit. This is because, the domain being a relatively high resistance region, the current tries to go by the sides, as shown in fig. (13.10.1a), if such space is available, thus causing the domain current to decrease. This is what precisely necessary for the growth of the depletion, so the depletion region extends on the sides. The excess current on the sides accumulating particles on the left of the depletion thus causing the

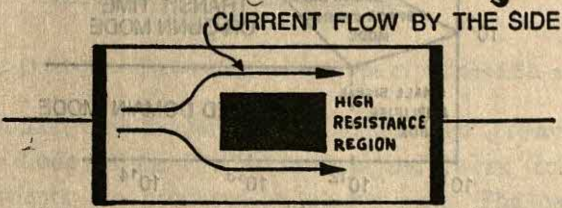


Fig. 13.10.1(a). Shows the device and the current pulse as the domain travels down the device.

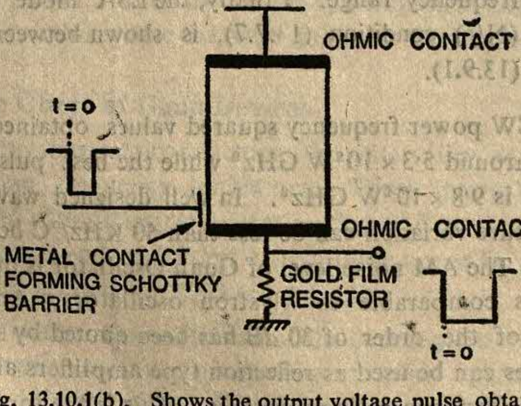


Fig. 13.10.1(b). Shows the output voltage pulse obtained by using a schottky barrier pulsing in a subcritically biased device.

growth of the accumulation region of the domain as well. The extension of the domain on the sides is also dependent on the interaction between two groups of particles (i) the slowly moving ones in the domain and (ii) those moving faster by the sides where the field is lower. The current density at the peak field position of the domain is given by $q\ N\ v_m$ where v_m is the velocity at E_m. Hence the current is

$$I = A(x)\, q\, N\, v_m \qquad \ldots \quad (13.10.1)$$

where the area A varies along the diode length. That is, any type of waveform of current can, on principle, be obtained in a load, connected in series with the TED device ; only by shaping the body of the diode suitably.

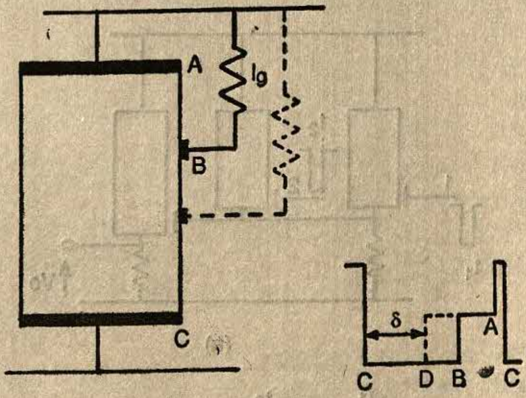


Fig. 13.10.1(c). Shows the cathode current as the domain passes the different points on the devices.

Second principle that is used for domain creation lies in the use of the diode with a Schottky barrier on its side near the cathode as shown in fig. (13.10.1b), where a subcritically biased (below V_T) diode has not have a domain propagating. But by application of a negative pulse, the depletion width is increased, thus causing momentary field concentration in the diode body adjacent to the depletion region and raising the field to above critical value there. This causes a domain to nucleate and propagate giving a current ; hence an output voltage waveform, as shown in the figure. The rising portion of the input pulse being positive going, does not affect the domain. In drawing the waveform, we have assumed instantaneous domain growth or collapse on reaching the anode. But this will modify only the sides of the waveform. The duration of the pulse is decided by the device length and is equal to the transit time of the pulse. Note that if the resistance was connected to the anode supply instead of the cathode, a positive going pulse would have been obtained at the anode to ground.

The third principle is that the domain having a depletion region in it acts as a high resistance and drops a large part of the applied voltage. Hence for a device with three ohmic contacts as shown in fig. (13.10.1c), biased above the threshold voltage, will have a domain nucleated at the cathode C when the current, out of the cathode C would decrease. When the domain is between B to C, large voltage drops between B to C ; hence less is available between A to B causing

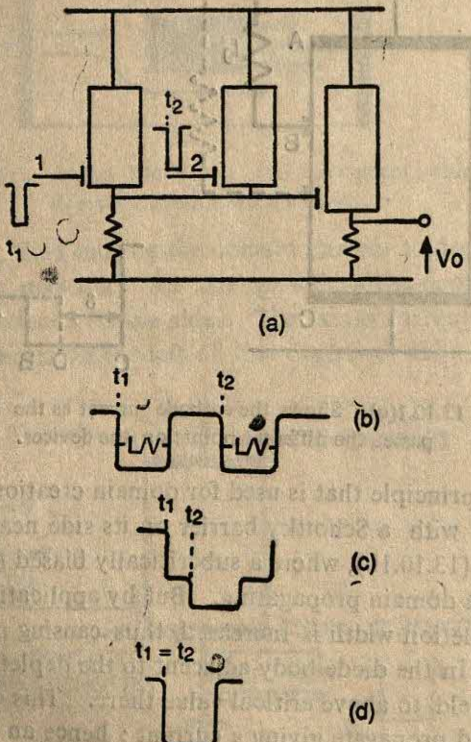


Fig. 13.10.2. (a) Shows the circuit, (b) shows the output voltage waveform if $(t_2 - t_1) > L/V$, (c) shows an AND simulation requiring a threshold, (d) shows an AND simulation if $t_1 = t_2$.

the current I_g to be small and the total cathode current constant. As soon as the domain has crossed the contact B, the portion AB has a large voltage drop across it. The current I_g rises while the domain current decided by the domain area remains same. So, the total cathode current rises at B. So long as the domain is in transit between B to A, the current again remains constant and as soon as the domain is delivered to the anode the current rises momentarily to fall again due to nucleation of another domain at cathode. As a matter of fact, this principle can be extended to obtain variable tapped delayed pulses in which the delay δ can be varied by tapping at different ohmic contacts along the diode body as shown by a dotted scheme in the figure.

Next, we examine a logic AND/OR circuit where the devices have Schottky gates and are connected as shown in fig. (13.10.2a). Application of a negative gate pulse at Schottky gate terminal 1 at time t_1 and at gate 2 at time t_2 will give voltage waveforms at the output as shown in fig. (13.10.2b) if $(t_2 - t_1)$ is greater than the transit time (L/v) of the first two devices on the left. This provides the simulation of OR. If $(t_2 - t_1)$ is less than (L/v), then the waveform is expected to be of the form shown in fig. (13.10.2c) which is an AND requiring a threshold*. If, on the other hand, $(t_1 - t_2) = 0$, the waveform is also shown in fig. (13.10.2d) *i.e.*, an AND operation.

NOT gate operation and INHIBIT can be realised by the circuit shown in fig. (13.10.3a) where the total bias voltage is subcritical. A negative going input pulse at input 1 at time t_1 and another at terminal 2 at time t_2 will give the output as shown in (13.10.3b), provided $(t_2 - t_1)$ is greater than (L/v). Note that the pulse at t_2 gives a "not" operation. If, of course, $(t_2 - t_1)$ is less than (L/v), the output would be given by fig. (13.10.3c), since once a domain has nucleated in the upper diode ; the total voltage across the lower diode is so small that the gate pulse is unable to bring it above V_T. Hence an inhibit operation is obtained.

---

* Note that such waveforms may be obtainable in cases where the NL product is on the lower side. In a subcritically doped diode, the domain may not mature during the transit time. The second pulse only enhances the maturity.

Similarly, in (13.10.3d), the pulse at t_1 is inhibited that occurred after time t_2.

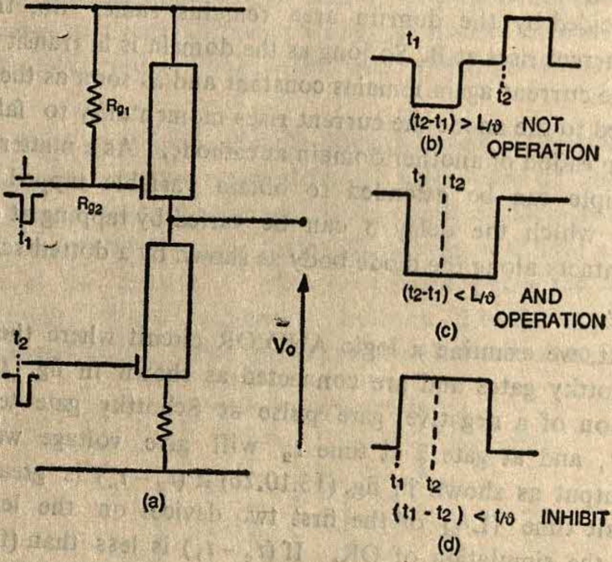


Fig. 13.10.3. (*a*) Shows the TED circuit for the operations mentioned in (b) NOT (c) AND (d) INHIBIT.

Next, we consider a memory circuit, as shown in fig. (13.10.4a). The voltage across each diode is less than V_T. Application of a negative going set pulse produces a voltage waveform at the output of the first device as shown in the diagram. When the domain traverses the notch region, a sharp negative pulse peak is obtained in the waveform. At this negative peak the second diode domain is nucleated when a similar waveform is obtained at point 2 which, in its turn, being coupled to the first diode input, turns the first diode on in time. Thus a set pulse keeps a bit stored in. A positive going reset pulse of suitable amplitude can inhibit the first diode from domain nucleation when the memory would be cleared. Another memory circuit is shown in fig. (13.10.4b), where the device is also biased subcritically, but the load has a capacitance across it. A negative going set pulse nucleates a domain at the cathode end. The device resistance increases due to the presence of the domain; hence the device current drops. The condenser discharges through the cathode resistance to a voltage at A, as shown in the fig. (13.10.4c). When the domain has reached the anode, the device current rises but the output voltage from A rises slower than the current. But during this time diode bias is greater than V_T, so another domain nucleates at B when the output voltage again falls. This continues until a positive going reset pulse forward biases the Schottky to inhibit the domain formation.

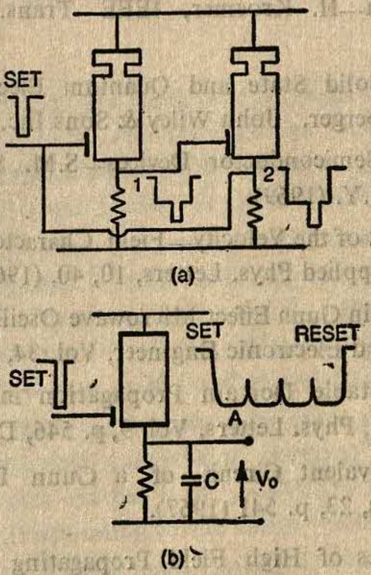


Fig. 13.10.4. (a) Shows an astable type of TED memory circuit, (b) shows another TED memory dependent on RC product for its operations.

References

1. The Possibility of Negative Resistance Effects in Semiconductors —B.K. Ridley and T.B. Watkins. Proc. Phys. Soc. (London), 78, 293, (1961).
2. 'Transferred Electron Amplifiers and Oscillators'—C. Hilsum Proc. IRE, 50, 185 (1962).
3. Microwave Oscillation of Current in III-V Semiconductors— J.B. Gunn, Solid State Comm. 1, 88, (1963).

4. (*a*) Theory of the Gunn Effect—H. Kroemer, Proc. IEEE, 52, 1736 (1964).
   (*b*) The Gunn Effect under Imperfect Cathode Boundary Condition—H. Kroemer, IEEE Trans., ED-15, p. 819 (1968).
5. Topics of Solid State and Quantum Electronics—Edited by W.D. Hershberger. John Wiley & Sons Inc. (1972).
6. Physics of Semiconductor Devices—S.M. Sze, John Wiley & Sons, Inc., N.Y. (1969).
7. Measurement of the Velocity—Field Characteristics of Gallium Arsenide—Applied Phys. Letters, 10, 40, (1967).
8. Mechanisms in Gunn Effect Microwave Oscillators—J.E. Carrol, The Radio and Electronic Engineer, Vol, 34, July (1967).
9. Theory of Stable Domain Propagation in the Gunn Effects. P.N. Butcher, Phys. Letters, Vol. 9, p. 546, Dec. 15, 1965.
10. On the Equivalent Circuit of a Gunn Diode—W. Heinle, J. Electronics, 23, p. 541 (1967).
11. The Dynamics of High Field Propagating Domains in Bulk Semiconductors—K. Kurokawa, Bell Syst. Tech. J. 46, 2235 (1967).
12. Hot Electron Microwave Generators—John E. Carroll, Edward Arnold (Publications) Ltd., London (1970).
13. LSA Oscillator Diode Theory—J.A. Copeland, J. Appl. Phys., 38, 3096, 1967.
14. Doping Uniformity and Geometry of LSA Oscillator Diodes —J.A. Copeland. IEE Trans. ED, ED.-14, 497 (1967).
15. LSA Operation of Large Volume Bulk GaAs Samples—Kennedy, W.K., Eastern L.F. and Gilbert R.J.—IEEE Trans. ED, ED-14, 500 (1967).
16. LSA Operation of Bulk *n*-GaAs Diodes—Copeland, J.A. and Spiwak, R.R.—Intern. Solid State Circuits Conf. Philadelphia (1967).
17. An Experimental Study of High Peak Power Gunn Effect Oscillators—Heeks J.S., Au, H.K. and Woode A.D.—7th Int.

Conference Microwave Optical Generation and Amplification, Hamburg (1968).

18. Information Prccessing using GaAs—RCA Review, March, 1973, Vol. 34, No. 1, p. 152-163.

Problems

1. If in fig. (13.7), the static and dynamic cha racteristics of the device is as in fig. (13.7.4) draw the voltage waveform V and form the expression for this voltage to ground. Assume the device is of length L and has unifrom cross-section.
2. Draw the current pulse waveform in quenched domain mode. Assume domain is formed at $E_p - E_r = 3{\cdot}2 \times 10^3$ KV/cm and quenching occurs at $E = E_s = 0{\cdot}5\ E_p$ instantaneously. Assume a diode 100 microns × 10 sq. microns and the bias voltage is 8·2 volts, with $v_p = 2 \times 10^7$ cm/sec and diode doping 10^{15}/cc.
3. Assuming that the domain acts like a constant capacitance C_a, establish that the domain charging current i_a is related to the circuit current i by $i_a = dQ/dt = i - I_s$ where I_s is the current corresponding to the valley velocity.
4. In quenched domain mode operation with the load as a series resonant circuit, assume the domain capacitance C_a constant and neglect C_0. Draw the domain charge versus time. Note that it should be part of sinusoid since C_a forms part of the resonant circuit. Hence draw the domain current. So, draw the circuit current waveform. Note that the maximum value of circuit current can be given by $i_{\max}\ R_0 = E_p.L$ [and the domain is discharged completely *i.e.*, $i_a = 0$ at $i = I_s$. Draw the device voltage *viz.*, applied voltage minus the circuit voltage. Show that the domain discharge requires a delay.
5. Show that in LSA mode the [gain exponent can be expressed as $\omega.\delta_1 = 2\omega_{cn}\,\theta_n$ while the loss exponent as $\omega\delta_2 = 2\omega_c\theta_p$ where $\omega_{on}\ \varepsilon\varepsilon_0 = qN\,\bar{\mu}_n$. Since for LSA, $G = G_n.G_p = \exp(-\delta)$ and $\delta = 1$ is required, show that $2(\omega_{cn}\,\theta_n - \omega_c\,\theta_p) = \omega$, find its value if $\mu_0 = 3\bar{\mu}_n$. Note that this gives a possible range of conduc tion angles for a certain device working in LSA mode.
6. Show that in fig. (13.7.4) $b_1 \cos\theta_p = (b_1 - 1)$ while for $\mu_0 = 3\bar{\mu}_n$ $[b_1 \cos\theta_p - \cos(\theta_p + \theta_n)] = 3(1 - a)$.

7. Draw the current waveform from piece-wise linear model as discussed in the text for LSA mode. Find the efficiency using the current and voltage waveforms in this mode.

8. Show that the power output per electron normalised to threshold power ($q\ E_p . v_p$) can be written as $\left[\int_0^{2\pi} v\, b_1 \cos\theta\, d\theta/2\pi v_p\right]$ in LSA mode where the symbols have their usual meaning.

9. Since $dE = E_p b_1 \sin\theta\, d\theta$ and the total power P_0 for all the involved electrons is in the same proportion to the threshold power $P = I_p E_p . L$, show from the result of problem (13.8) by first integration by parts and substituting properly that,

$$\left(\frac{P_0}{p}\right) = -\int_0^{2\pi} b_1^2 \sin^2\theta \left(\frac{dv}{dE}\right) d\theta/2\pi\ \mu_0.$$

10. Using the results of problem (13.9) with $\delta = 0$ and $\mu_0 = 3\bar{\mu}_n$ of problem (13.5), show that $P_1/P - b_1^2\ [4 \sin 2\theta_n - \sin 8\ \theta_n/12\pi.]$, P_1 being the fundamental R.f. power.

Chapter 14

MESFET

It has been noted in Chapters (9) and (11) that the figure of merit of JFET and MOSFET can be increased by decreasing the channel length. This led to the investigations on short channel length JFET and IGFET. It has also been shown that in special structures like the D-MOS and V-MOS, certain technological advantages are obtainable in increasing the power and frequency performance through increase of channel width and decrease of channel length respectively. In article (9·3), it has been shown that the material conductivity also plays a prominent role in frequency and power performance of JFET. Low resistivity channel is expected to decrease channel noise.

In both the JFET and IGFET the minimisation of input and output capacitances are necessary to improve the frequency performance. It should be noted that the gate bonding pad to substrate capacitance adds to input side as also to the output side. A high resistivity epitaxial buffer layer can be used to isolate the current carrying (active) layer from the substrate thus minimising these capacitances. This, in turn, is expeeted to minimise noise as well.

On the other hand, a channel of a high mobility material would always give a higher g_m (cf. equation 9.1.13 and 11.2.6). In the saturation region of FET the field in the pinch-off region is normally in the saturation velocity range. Hence a material with higher saturation velocity is expected to give a lower transit time. For high power, a high breakdown voltage of the material that is a higher E_g material will be useful. In addition, the source drain dopings can be made N^+ on N that would minimise parasitic contact resistance as also will increase drain source breakdown voltage. To lower gate metal resistance under the long gates in multigate high power devices, short gate branches can be used. This would decrease the noise in turn. For better heat sink, flip chip (planar construction inverted) mounting can be used, thus decreasing the thermal resistance and minimising ground source lead inductance.

IGFET, of course, has advantages over JFET in that for N-channel MOST the gate voltage can swing from depletion mode to enhancement mode without any gate bias requirement. Moreover, in the active region of an enhancement mode MOST, the input capacitance and transconductance are almost independent of gate voltage and the output capacitance to drain voltage. This gives very linear (class A) power amplification with low amplitude and phase distortion. Unfortunately, practical short channel ($<$ 1 micron) IGFETs, so far studied, are mostly of silicon material and with GaAs there has only been limited success (cf. ref. 1).

The above discussion points to the prospective member as JFET for high frequency and high power purposes, while Si and GaAs for which the technologies are well established, are the only two materials, with which large number of devices have been constructed and studied. GaAs has six times larger mobility and two times peak drift velocity than Si ; hence whih GaAs higher g_m and less transit time is expected. It has higher band gap ; hence has higher breakdown voltage than Si, a higher E_g that can give much higher resistivity (greater than 10^7 ohm-cm) substrate. In contrast to Si, it has self-ballasting capability due to increased resistance before burn out. In addition, polycrystalline GaAs passivation is possible, can give into self aligning ; thus enabling reduction of input and output capacitances.

Schottky barrier on the other-hand has higher input impedance than a *p-n* junction. The above ideas led to the invention of short channel Metal Semiconductor Field Effect Transistors (MESFET s) using GaAs. These are basically majority carrier devices made of N-type channel region, the cross-section of which is controlled by a reversed bias Schottky metal gate. JFETs has not yet been able to compete with the MESFETs mainly because of technological limitations of fabrication of channel lengths shorter (less than 0·5 microns) by diffusion. For a good understanding of the performance capabilities of the GaAs MESFETs and comparison with the other members mentioned above, the reader should consult ref. (1). In the next article we discuss the operating principle of submicron length GaAs MESFET.

14.1 Principle of Operation

It has been mentioned that the MESFET is a Schottky barrier controlled metal semiconductor junction FET. For an understanding

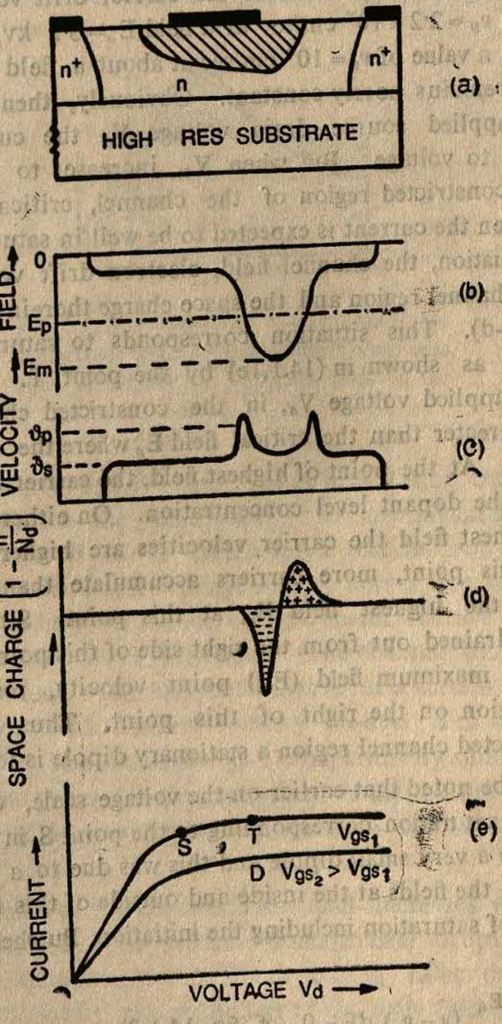


Fig. 14.1.1. Shows (a) the depletion region and channel in the MESFET, (b) field in the channel region, (c) the corresponding velocity, (d) the space-charge corresponding to the point 'T' in the I-V characteristics in (e). [Ref. C.A. Liechti © Copyright 1976 IEEE, Trans. MTT-24 by permission]

of the nature of charge, field or velocity veriation within the channel region, an N-doped GaAs MESFET is shown in fig. (14.1.1a). It is to be remembered that GaAs has a velocity field characteristics as given in fig. (13.1.1). That is the carrier drift velocity reaches a peak value $v_p = 2{\cdot}2 \times 10^7$ cm/sec at a field $E_p = 3{\cdot}2$ kv/cm, decreases thereafter to a value of $v_s = 10^7$ cm/sec at about a field $E_q - 20$ kv/cm and then it remains nearly constant. Obviously, then at very low values of applied source drain voltage V_d the current will be proportional to voltage. But when V_d, increases to a value such that in the constricted region of the channel, critical field E_p is exceeded, then the current is expected to be well in saturation region. For this situation, the channel field, electron drift velocity in the constricted channel region and the space charge therein are shown in figs. (14.1.1b-d). This situation corresponds to saturated current, well inside, as shown in (14.1.1e) by the point T. That is, for appreciable applied voltage V_d, in the constricted channel region, the field is greater than the critical field E_p where the carriers have low velocity. At the point of highest field, the carrier concentration is equal to the dopant level concentration. On either side of the point of highest field the carrier velocities are higher. Hence, on the left of this point, more carriers accumulate than that can be drained by the highest field E_m at this point. Similarly, more carriers are drained out from the right side of this point than those given by the maximum field (E_m) point velocity, resulting in a depletion region on the right of this point. Thus we see that, in the constricted channel region a stationary dipole is formed.

It is to be noted that earlier on the voltage scale, when current reached just saturation corresponding to the point S in fig (14.1.1e), there was also a very small dipole and this was due to a small differrence between the fields at the inside and outside of this region. Since in all cases of saturation including the initiation, Butchers equal area rule

$$\int_{E_0}^{E_m} (v - v_n)\, dE = 0, \text{ (cf. fig. 14.1.2),}$$

holds, this obviously gives the maximum field E_m for stable stationary domain formation. From this equal area rule, we can conclude that, starting from the low field, the saturation was reached when the highest field just exceeded the peak field E_p, and E_0 – the outside domain field was nearly equal to the maximum field E_m. This indicates that the depletion region under the Schottky gate in GaAs has a flatter boundary than in a Si MESFET which obviously would be of the type as discussed in JFET (cf. Chapter 9). Since nearly equal field implies nearly equal velocity, hence nearly equal channel thickness is required for the same current to flow just at the onset of saturation. But it is to be remembered that as soon as the little domain formed, most of the voltage drops across the domain, the channel widens a little on the left of the domain because of the reduced Schottky gate bias at this region of the channel, thus letting the same current to flow. This happens because for a certain current the channel thickness and channel velocity product is to be constant; while high mobility and short gate give comparatively smaller voltage change across the channel region compared to that occurs with silicon due to field readjustment under the gate on the left of the space charge.

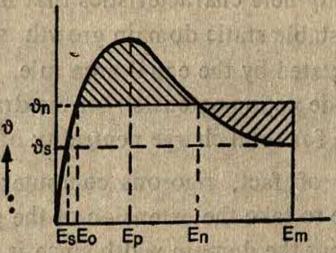


Fig. 14.1.2. Shows the velocity field characteristics of GaAs and Butchers equal area rule as related to different fields in the device.

With increase of V_a, the maximum field increases which drops more voltage, thus E_o, the outside domain field drops further as required by the equal area rule. The drop in E_0 value decreases the reverse bias on Schottky gate at the left of the high field region thus widening the channel thickness in this region as required for increase of current in the saturation regions with increased source drain voltage V_a. Thus in the saturation region, the domain formed in the constricted channel dictates the current voltage characteristics.

On the other hand, in this region the channel depletion on the right of the maximum field must extend well out of the gate edge at the drain end. This is because we have noted (cf. Chapter 9) that even if the material was Si instead, under current saturation, there formed a depletion region on the right of the maximum field. GaAs has a typical velocity field characteristics that makes this effect more pronounced. The stable static domain growth so the accumulation of electrons is dictated by the equal area rule. But the doping level is fixed ; hence the depletion extension on the drain side is the only variable to compete for total charge neutrality of the domain.

As a matter of fact, rigorous computer calculations indicate that it is reasonable to take the extension of the gate depletion beyond the gate and by about the domain width since it is this portion of the space charge that compresses the domain to the required size so that the domain produces the correct field to avoid its propagation. When the gate becomes more negative the Schottky gate reverse bias increa-

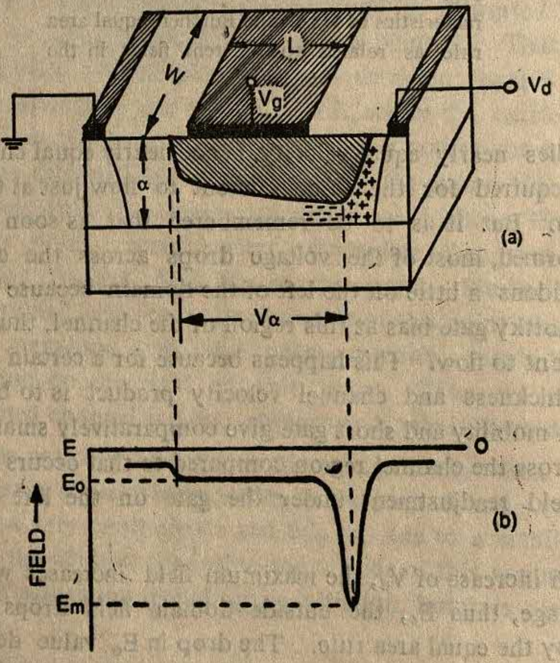


Fig. 14.1.3. Shows in (a) the depletion region and space charges and (b) the field along the channel region. [Ref. M.S. Shur, by permission. Copyright © 1978 IEEE.]

ses all through the channel length. This decreases the channel thickness further. In the space charge regions, the domain drops more voltage in addition, leading to an overall current reduction and an increase of the depletion width on the drain end. This characteristic is shown in fig. (14.1.1e) as curve D.

From the above discussion with fig. (14.1.3), we can easily find out the current voltage equation of the GaAs MESFET. So long the conditions in the channel is in ohmic low field region, the current flow can be described as for JFET before saturation. That is if the voltage drop across the gate region along the channel (V_a), is less than the domain sustaining voltage

$$V_s = E_s L \qquad \ldots \quad (14.1.1)$$

the channel current voltage relation can be written from equation (9.1.11) as

$$I_c = g_0[V_a - \tfrac{2}{3}\{(V_a + V_b + V_g)^{\frac{3}{2}} - (V_b + V_c)^{\frac{3}{2}}\}/(V_p)^{\frac{1}{2}}] \quad \cdots \quad (14.1.2)$$

where $$g_0 = (a\, W \mu q N)/L \qquad \ldots \quad (14.1.3)$$

is the open channel conductance, and

$$V_p = (qN\, a^2/2\varepsilon) \qquad \cdots \quad (14.1.4)$$

is the so called channel pinch-off voltage. Normally $V_a \leqslant (V_b + V_g)$, when equation (14.1.2) can be approximated to

$$I_c = g_a V_a \qquad \cdots \quad (14.1.5)$$

where $$g_a = g_0[1 - \sqrt{(V_b + V_g)/V_p}\,] \qquad \cdots \quad (14.1.6)$$

The saturation current is given by the equations (14.1.5) and (14.1.6) when $V_a = E_0 L$, that is the voltage required for domain formation by satisfying Butcher's equal area rule. But as discussed previously this voltage settles down to the domain sustaining voltage $V_s = E_s L$ due to the increase of the depletion beyond the gate. Hence the saturation current becomes

$$I_s = g_a V_s \qquad \cdots \quad (14.1.7)$$

The transconductance in the saturation region is obtained from equation (14.1.2) as

$$g_m = \left.\frac{\delta I_c}{\delta V_g}\right|_{V_a = V_s} = g_0[(V_s + V_b + V_g)^{\frac{1}{2}} - (V_b + V_g)^{\frac{1}{2}}]/V_p^{\frac{1}{2}} \qquad \cdots \quad (14.1.8)$$

Since in short channel MESFETs, $V_s \ll (V_b + V_g)$, equation (14.1.8) can be approximated using equations (14.1.3), (14.1.4), (14.1.1) and $v_s = \mu E_s$ where E_s is the domain sustaining field nearly equal to E_s of figure (14.1.2), as

$$g_m = W\, v_s [qN\varepsilon/2(V_b + V_g)]^{\frac{1}{2}} \qquad \text{... (14.1.9)}$$

The depletion charge under the gate Q_g, below current saturation, can be obtained from $t(x)$, the depletion width as function of x, as

$$Q_g = qNW \int_0^L t(x)\, dx \qquad \text{... (14.1.10)}$$

where for $t(x)$ we have like equation (9.12)

$$t(x) = a[(V(x) + V_b + V_g)/V_p]^{\frac{1}{2}} \qquad \text{... (14.1.11)}$$

A first order approximation to the channel field is adequate (cf. ref. 2), where the channel field is assumed to be constant $E_0 = E_a = (V_a/L)$ that is

$$V(x) = E_a . x \qquad \text{... (14.1.12)}$$

Suitable high doping combined with high mobility in MESFET allows this approximation to hold nicely (cf. ref. 3). Performing the integration of equation (14.1.10) with (14.1.11) and (14.1.12) and substitution from equation (14.1.4) for V_p with $V_a = E_a . L$, one has

$$Q_g = 2\sqrt{2}\ WL(\varepsilon\, qN)^{\frac{1}{2}}[(V_a + V_b + V_g)^{\frac{3}{2}} - (V_b + V_g)^{\frac{3}{2}}]/3V_a \qquad \text{... (14.1.13)}$$

which, for V_a very much less than $(V_b + V_g)$, can be approximated as

$$Q_g = \sqrt{2}\ WL\,[\varepsilon\, qN(V_b + V_g)]^{\frac{1}{2}} \qquad \text{... (14.1.14)}$$

In this triode region the gate drain and source gate capacitances can be easily found from equation (14.1.13). For the first one

$$C_{dgo} = \left.\frac{dQ_g}{dV_g}\right|_{V_a = \text{const.}} \qquad \text{... (14.1.15)}$$

while for the gate source capacitance

$$C_{gso} = \left.\frac{dQ_g}{dV_g}\right|_{V_a + V_g = \text{const.}} \qquad \text{... (14.1.16)}$$

These two capacitances found out with equation (14.1.13), can be written in approximate form as

$$C_{ago} = C_{gso} = WL\,[(\varepsilon\, qN/2(V_b + V_g)]^{\frac{1}{2}} \quad \cdots \quad (14.1.17)$$

The figure of merit in the triode region can be found using equations (14.1.9) and (14.1.17) as

$$\omega_T = (g_m/C_{gso}) = (v_s/L) \quad \cdots \quad (14.1.18)$$

Note that since it is not V_d, the drain source voltage, but V_a, the voltage in the channel under the gate only, that appears in equation (14.1.2). Equation (14.1.13) is valid for the saturation region also. In this region, the voltage drop across the gate section of the channel region $V_a = V_s = E_s L$ where E_s is the domain sustaining field. In view of our earlier discussion, this is a situation when the device is well in saturation.

14.2 Equivalent Circuit

Since in a MESFET in saturation, the domain is formed near the drain end of the gate, the domain capacitance through Miller effect can largely determine the frequency and speed performance. Hence it would be appropriate to discuss the MESFET domain characteristics before embarking on the equivalent circuit. The domain voltage drop V_{dom} has been found earlier in Chapter (13) for an uniform cross-section. The problem here is a little different in that it is under the influence of the Schottky depletion region. The wedge shape of this depletion results in the highest domain field, a little removed on the drain side from the drain end edge under the gate electrode. This

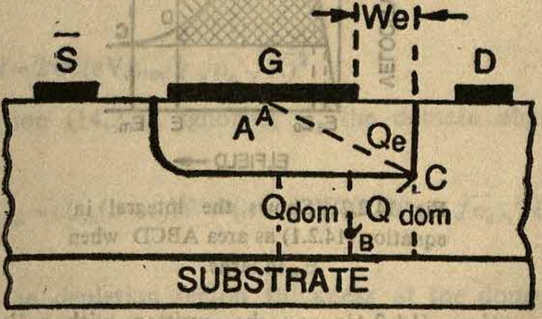


Fig. 14.2.1. Shows the ideal assumed boundaries of the differently charged regions in MESFET.

is a result of the field lines between the gate edge to the domain depletion. For ease of analysis, some simplifying assumptions are necessary regarding the share of the various charges to the different voltages. Fig. (14.2.1) shows these charges pictorially with their assumed ideal boundaries. In this figure, Q_{dom} is the assumed domain depletion charge, Q_g is the Schottky barrier depletion charge under the gate metal while Q_s is the shared field caused depletion charge beyond the gate metal. It is this charge that controls the domain size so that the domain is inhibited from moving as also from growing further than just necessary to let the required current to pass through.

Turning our attention to the stationary domain, in equation (13.5.5) if the velocity of the moving frame is set equal to the particle velocity external to the domain, then $J = qNu$ and the left hand term drops out. Integrating the resulting equation between E_0 to E and substituting ρ for accumulation of charge we have

$$\log\left[e^{p}/(1+p)\right] = \frac{\varepsilon}{DqN}\int_{E_0}^{E}\left[v(E) - u(E_0)\right]dE \quad \cdots \quad (14.2.1)$$

where
$$p = (\rho/qN) \quad \cdots \quad (14.2.2)$$

Equation (14.2.1) gives the Butcher's domain space charge equation.

In view of the identity (13.5.8) and fig. (14.2.2), the integral on the right of equation (14.2.1) is the area ABCD in the figure.

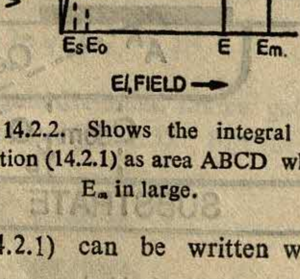


Fig. 14.2.2. Shows the integral in equation (14.2.1) as area ABCD when E_m in large.

So, the equation (14.2.1) can be written with a little error (cf. ref. 4) as

$$\log\left[e^{p}/(1+p)\right] = \varepsilon\mu\,(E_0 - E_s)(E_m - E)/qDN \quad \cdots \quad (14.2.3)$$

Since $E_m \gg E_s$, equation (14.2.3) with equation (13.5.8), on division gives

$$\log [e^{p}/(1+p)] = \frac{n_c}{N}(1 - E/E_m) \qquad \cdots \quad (14.2.4)$$

where

$$n_c = (\varepsilon \mu E_c^2/qD), \qquad \cdots \quad (14.2.5)$$

which for small values of p, that is small space charge, gives

$$p^2 = 2\frac{n_c}{N}\left(1 - E/E_m\right) \qquad \cdots \quad (14.2.6)$$

Small space charge case is seen to be (cf. ref. 4, 5) adequate in MESFET approximate description.

Domain voltage can be found from (cf. ref. 6)

$$V_{dom} = \int_{E_s}^{E_m} (E - E_s)\, dx \qquad \cdots \quad (14.2.7)$$

Since $E_s \leqslant E_m$ and $\rho = \frac{\delta E}{\delta x}$ while the space charge given by equation (14.2.6) is symmetrical, we have from equation (14.2.7), normalising with E_m

$$V_{dom} = 2\varepsilon (E_m - E_s)^2 \int_0^{E_m} \frac{E}{\rho E_m} \cdot d\left(\frac{E}{E_m}\right) \qquad \cdots \quad (14.2.8)$$

Substituting for ρ from equation (14.2.6) and integrating, we have

$$V_{dom} = 1{\cdot}89\varepsilon (E_m - E_s)^2/q\sqrt{n_c}\sqrt{N} \qquad \cdots \quad (14.2.9)$$

Symmetrical domain width d, obtained by integrating Poisson's equation with ρ from eqn. (14.2.6), can be written with equation (14.2.9.) as

$$d = 2{\cdot}06(\varepsilon V_{dom}/q\sqrt{n_c}\sqrt{N})^{\frac{1}{2}} \qquad \cdots \quad (14.2.10)$$

Using equation (14.2.9), ignoring E_s the domain charge can be obtained as

$$Q_d = \varepsilon E_m(a - t_o)W = (a - t_o)W(0{\cdot}729\, q\varepsilon \sqrt{n_c}\sqrt{N}\, V_{dom})^{\frac{1}{2}} \qquad \cdots \quad (14.2.11)$$

where t_0 is the depletion region thickness at the domain location and is given by equation (14.1.11) when $V_s \leqslant (V_g + V_b)$.

The domain resistance can be obtained with equation (14.1.7) as

$$R_{dom} = \frac{\delta V_{dom}}{\delta I_s}\bigg|_{V_g\ \text{const.}} = \frac{\delta V_{dom}}{g_d L\ \ \delta E_0}\bigg|_{V_g\ \text{const.}} \quad \ldots \quad (14.2.12)$$

since it is E_0, the external domain field that readjusts to changes to settle down to E_s, using the identity (13.5.8) in its approximate form $E_o^2 = E_m(E_0 - E_s)$, noting that $\delta(E_0 - E_s) = \delta E_0$ and eliminating E_m using (14.2.9), we have from equation (14.2.12)

$$R_{dom} = -(\sqrt{\varepsilon}\, V_{dom}^{\frac{3}{2}})/q^{\frac{3}{2}} DW(a - t_0)(n_c N)^{\frac{3}{4}} \quad \cdots \quad (14.2.13)$$

In fig. (14.2.3a), the different distributed circuit elements are lumped at locations that enable visualisation of their origin. The equivalent circuit of fig. (14.2.3b) can be compared with it. The series elements R_{gs} and C_{gs}, C_{gd} and R_{gd} as also the substrate resistance R_{sub} are distributed elements, lumped between intrinsic source S′ and intrinsic drain D′ points. Normally, $R_{sub} \gg |R_{dom}|$ while for short channel MESFETs (of the order of micron) $(1/g_d)$ the

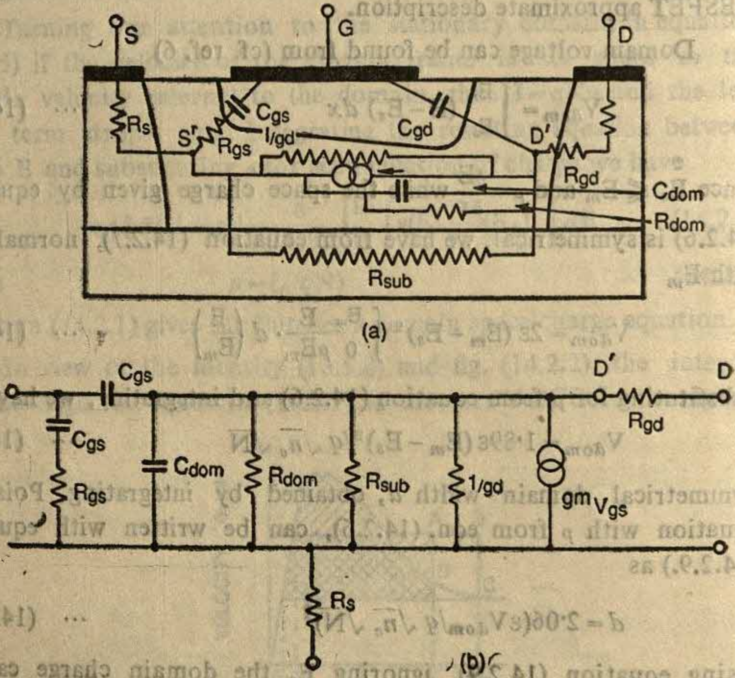


Fig. 14.2.3. Shows in (a) the locations when the elements are lumped. (b) the corresponding equivalent circuit (Ref. 6, by M. S. Shur)

drain resistance tends to smaller value with decrease of channel length. So, the combined effect of the three parallel resistance tends to a small positive resistance. R_s is the source series resistance that effectively reduces g_m according to g_m effective $= g_m/(1 + R_s g_m)$.

The different capacitances that appear in the equivalent circuit can be defined for the saturation region of operation as follows :

$$C_{gs} = \left.\frac{d(Q_g + Q_e)}{dV_g}\right)_{(V_d + V_g)\ \text{const.}} \qquad \cdots \quad (14.2.14)$$

$$C_{dg} = \left.\frac{d(Q_g + Q_e)}{dV_d}\right)_{V_g\ \text{const.}} \qquad \cdots \quad (14.2.15)$$

$$\text{and } C_{dom} = \left.\frac{dQ_d}{dV_{dom}}\right)_{V_g\ \text{const.}} \qquad \cdots \quad (14.2.16)$$

where Q_e is the depletion field shared charge and can be found by assuming that the depletion region extension beyond the gate is proportional to the domain width

$$W_e = \alpha d \qquad \cdots \quad (14.2.17)$$

Space charge profiles, generated by rigorous computer calculations, justify this assumption with α of the order of unity. Presumably, this originates from the property of the domain to spread up to fill the whole cross-section (cf. art. 13.10). The charge Q_e therefore can be written as

$$Q_e = qN\, i_0\, d\alpha W \qquad \cdots \quad (14.2.18)$$

From equation (14.2.15), C_{gd} can be found out easily by noting that V_g is constant while V_s is very much less than either V_{dom} or $(V_b + V_g)$ or both in the region when we have the source drain voltage $V_d \simeq V_{dom}$ and using equation (14.1.13) for saturation region with constant V_g which is invariant of V_{dom}.

$$\text{Hence} \quad C_{gd} = \left.\frac{dQ_e}{d\,V_{dom}}\right)_{V_g} \qquad \cdots \quad (14.2.19)$$

From equation (14.2.18) with (14.2.10), (14.1.11) and (14.1.14), we have

$$C_{gs} = 1{\cdot}46\alpha W[N(V_b + V_g)/V_d n_e]^{\frac{1}{2}} \qquad \cdots \quad (14.2.20)$$

C_{dom} can be found from equation (14.2.16) with equation (14.2.11) as

$$C_{dom} = 0{\cdot}364\, W(a - t_0)(q\varepsilon \sqrt{n_c} \sqrt{N}/V_{dom})^{\frac{1}{2}} \quad \dots \quad (14.2.21)$$

while

$$\begin{aligned} C_{gs} &= \left[\frac{\delta Q_g}{\delta V_g} + \frac{\delta Q_c}{\delta V_g}\right]_{(V_d+V_g)} \\ &- \frac{\delta Q_g}{\delta V_g}\bigg|_{(V_d+V_g)} + qN_aW\left[d \cdot \frac{\delta t_0}{\delta V_g} - t_0 \cdot \frac{\delta d}{\delta V_g}\right] \\ &= WL[\varepsilon q N/2(V_b + V_g)]^{\frac{1}{2}} + C_{gd}\left(\frac{V_d}{V_b + V_g} - 1\right)\left(\frac{\sqrt{n_c}}{\sqrt{N}}\right)^{\frac{1}{2}} \end{aligned}$$

since $(V_g + V_d) =$ constant implies $\delta V_g = -\delta V_d = -\delta V_{dom}$. The student should consult reference (3,4) for experimental verification and limitations of the above parameters of the equivalent circuit.

14.3 Design Considerations

In this article, we discuss the design criteria of MESFET assuming that a stationary large size fully developed domain is present in it along the lines described in reference (6). From the above discussions (cf. fig. 14.2.1—distance CA and CB) where Q_e terminates on the negative gate ; hence the depletion field near the middle of the gate is expected to be higher than the domain field since for the same exposed charge density, the depletion width of $(Q_g + Q_e)$ is greater than that of Q_{dom}. While experiments show that light emission at the breakdown of MESFET occurs between the gate and drain. This indicates that the domain breakdown may be the cause of device failure. This discrepancy can be explained by the results of computer calculations using field dependence of ionisation coefficient (cf. ref. 5). The field at which the breakdown occurs is considerably higher in the totally depleted narrow region under the gate compared to the breakdown field within the quasi-neutral (cf. equation 14.2.6) relatively wide high field domain. This is because the ionisation integral requires for a certain field, (hence α) a certain space. Hence restriction in space requires high α ; hence high field. Typical values of gate breakdown field E_{gb} and domain breakdown field E_{db}, used by these authors are 300 kv/cm and 200 kv/cm. The other physical parameters as taken by these authors are $\varepsilon = 12{\cdot}9\, \varepsilon_0$, (Sze quotes 10·9), $\mu_0 =$

4500 cm^2/V sec, domain sustaining Field $E_s = 1{\cdot}89$ kv/cm, $n_{cr} = 1{\cdot}6 \times 10^{15}/cm^3$ and $v_s = 0{\cdot}85 \times 10^7$ cm/sec.

A few design criteria may be set from general considerations as follows (cf. ref. 6) :

(1) The source drain voltage V_d is nearly equal to the domain voltage. So, it should not exceed the domain breakdown voltage, BV_d. That is the domain maximum field should not exceed domain breakdown field E_{db}.

(2) Considering biasing on the transfer characteristic, for class B operation, maximum gate swing can be equal to twice the pinch-off voltage, since forward biasing of Schottky gate is to be avoided. At the negative peak of this swing breakdown under the gate should not occur. Hence twice the pinch-off voltage should be less than the gate breakdown voltage BV_g.

(3) The operating frequency will be less than f_T (cf. equation 14.1.18). But the actual device f_T is expected to be lower than the intrinsic f_T because of the presence of the extrinsic parameters.

(4) The gate width W should not exceed a certain fraction of the wave-length in the medium. Otherwise, quite a few additional effects are expected to interfere in the performance.

(5) The thickness of the open channel 'a' should not exceed half of the gate length. Since the effective gate length $L_{eff} = (a + L)$ would otherwise increase considerably worsening the frequency response. Computer studies (cf. ref. 1) show that for submicron MESFET, decrease of gate length increases f_T while the output resistance decreases. The limit is reached for $(L/a) = 1$. Therefore, a factor of two safety is reasonable.

Since E_m is very much greater than E_s, equation (14.2.9) gives the maximum domain field, which should not exceed the domain breakdown field E_{db} mentioned in criteria (1) above. Hence, we have from equation (14.2.9)

$$V_d \leqslant 1{\cdot}89\, E_{db}{}^2/q \sqrt{n_c} \sqrt{N} = BV_d \qquad \cdots \quad (14.3.1)$$

On the other hand, to house the fully developed domain the drain gate spacing d_{dg} should be at least greater than the domain width. That is from equation (14.2.9) and (14.2.10)

$$d_{dg} \geqslant d = 2{\cdot}83(\varepsilon\, E_{db}/q \sqrt{n_e} \sqrt{N}) \qquad \cdots \quad (1\ .3.2)$$

From the restriction (3) above and equation (14.1.18), we can write

$$L = (\theta\ v_s/\pi\ f_T) \quad \cdots \quad (14.3.3)$$

where θ is a fraction.

Equation (14.3.3) with the restriction (5) above gives

$$a = (L/2) = (\theta\ v_s/2\pi f_T) \quad \cdots \quad (14.3.4)$$

While restriction (4) gives

$$W = K\lambda = (Kc/f_T\sqrt{\varepsilon_s}) \quad \cdots \quad (14.3.5)$$

From restriction (2), we can write

$$2V_p \leqslant BV_g = (\varepsilon\ E_{gb}{}^2/2qN) \quad \cdots \quad (14.3.6)$$

since a step junction approximation can be made of the gate depletion. Equation (14.3.6) with equation (14.1.4) for V_p gives

$$Na \leqslant \varepsilon\ E_{gb}/q\sqrt{2} \quad \cdots \quad (14.3.7)$$

which with (14.3.4) gives

$$N < f_T(\sqrt{2}\ \pi\ \varepsilon\ E_{gb})/\theta q v_s \quad \cdots \quad (14.3.8)$$

Substituting N from equation (14.3.8) in equation (14.3.2) above

$$d_{dg} = 1{\cdot}34\ (\theta\ \varepsilon v_s\ E_{db}{}^2/q\ n_c\ E_{gb}\ f_T)^{\frac{1}{2}} \quad \cdots \quad (14.3.9)$$

while with equation (14.3.1) gives

$$BV_d = 0{\cdot}9(\varepsilon\ v_s\ \theta\ E_{db}{}^4/qn_c\ E_{gb}\ f_T)^{\frac{1}{2}} \quad \cdots \quad (14.3.10)$$

The maximum power capability, on the other hand, would be limited by the maximum source drain voltage BV_d, the open channel current I_o and the class of amplifier operation. Hence

$$P \leqslant b\ I_o\ BV_d \quad \cdots \quad (14.3.11)$$

for Class A operation $(I_o/2\sqrt{2}).(BV_d/2\sqrt{2})$ hence $b = 1/8$ while for Class B operation $(I_o/\sqrt{2})(BV_d/2\sqrt{2})$ hence $b = 1/4$. We have for I_o

$$I_o = q\ N\ v_s a\ W \quad \cdots \quad (14.3.12)$$

Hence from equation (14.3.11) with the above equations 4, 5, 8, 10 and 12 we have

$$P = 0{\cdot}64bKc\ (\varepsilon^3 v_s{}^3\ \theta\ E_{gb}\ E_{db}{}^4/qn_c\ f_T{}^3\ \varepsilon_s)^{\frac{1}{2}} \quad \cdots \quad (14.3.13)$$

So, the power per unit gate width (P/W) can be found from (14.3.13) with (14.3.5). The diode impedance can be calculated as

$$Z=(BV_{d}/I_{o})=1{\cdot}27(\theta f_{T}/E_{gb}{}^{3}q\,\varepsilon_{o}\,n_{c}\,v_{s})^{\frac{1}{2}}\,(E_{bd}{}^{2}/Kc)\quad\cdots\quad(14.3.14)$$

Hence the power impedance product can be found from (14.3.13) and (14.3.14)

$$PZ=0{\cdot}81\,(b\,\theta\varepsilon\,v_{s}\,E_{db}{}^{4})/(qn_{c}\,E_{gb}\,f_{T})\quad\cdots\quad(14.3.15)$$

For MESFETs having gate length less than one micron, however, a different physical situation occurs and hence the above design procedure may not exactly be applicable for such MESFET.

For MESFETs less than one micron gate length, the carriers pass through the high field region in too short a time in which intervalley scattering is unable to cause velocity moderation effectively. That is the carriers do not settle down to the equilibrium velocity within their transit time through the stationary domain region. Hence, for such MESFET the equilibrium velocity field characteristics do not apply. That is the cariers getting accelerated can (cf. ref. 1) possess velocities very much higher than the peak velocity v_p. The velocity overshoot shortens the electron transit time through the high field region and shifts the accumulation layer into the drain gate space.

References

1. Microwave Field Effect Transistors—1976, Charles A. Liechti—IEEE Trans MTT, 24, June 1976, pp. 279.
2. Analytical Model of GaAs MESFET's—M.S.Shur, IEEE Trans ED, ED-25, No. 6, June 1978.
3. Small Signal Nonlinear Circuit Model of GaAs MESFET—M. S. Shur, Solid State Electronics, Vol. 22, p. 723.
4. Maximum Electric Field in High Field Domain—M. S. Shur., Electronics Letters, Vol. 14, No. 16, Aug. 1978.
5. Design Criteria for GaAs MESFETs Related to Stationary High Field Domains. L. F. Eastman, S. Tiwari and M. S. Shur. Solid State Electronics, Vol. 23, No. 4, p. 383.
6. Electronics Letters—M. S. Shur, Vol. 14, Issue 16, 1978.

Chapter 15

HETEROJUNCTION

A junction formed between two semiconductors of different energy gap is called an heterojunction. Shockley (1) in 1951 proposed that these junctions could be used as efficient emitters in bipolar transistor. Kroemer (2) in 1957 suggested their high injection efficiency possibility. In 1960 Anderson (3) first fabricated and studied these junctions. Since then these junctions have attracted the attention of physicists and engineers because of their having interesting useful properties arising out of discontinuities in the energy band at the junction. Heterojunctions are of two types. A junction formed of materials having similar type of conductivity (*n* or *p*) is called Isotype while a junction, formed of dissimilar conductivity type materials (*n-p* or *p-n*), is called an anisotype junction. There are controversies regarding the fitness of the convention to be used for the energy band model (4) regarding allocation of the discontinuity either in the so-called vacuum level or in the intrinsic level. We shall follow here Anderson's approach that assumes the vacuum level to be continuous.

15.1 Abrupt Anisotype Junction

Here we study the anisotype junction of two semiconductors having different band gaps (E_g), electron affinity (ψ), work function (ϕ) and dielectric permittivity (ε). The different physical constants for the two semiconductors when they are separately placed is shown in Fig. 15.1.1(a). When the junction is formed, the Fermi level equalises. The N-type material in this case spills electrons in P-type material. This causes the N-type material to have an electron depletion region and the P-side to obtain an electron accumulation region. This fact is depicted in Fig. 15.1.1(b) by appropriate bend in the bands at the junction region. Further, the energy band diagram must satisfy the continuity of the electric displacement vector perpendicular to the junction surface (interface charges being ignored). Starting from the Poisson's equations for each uniformly doped regions, one can easily show (assuming step junction

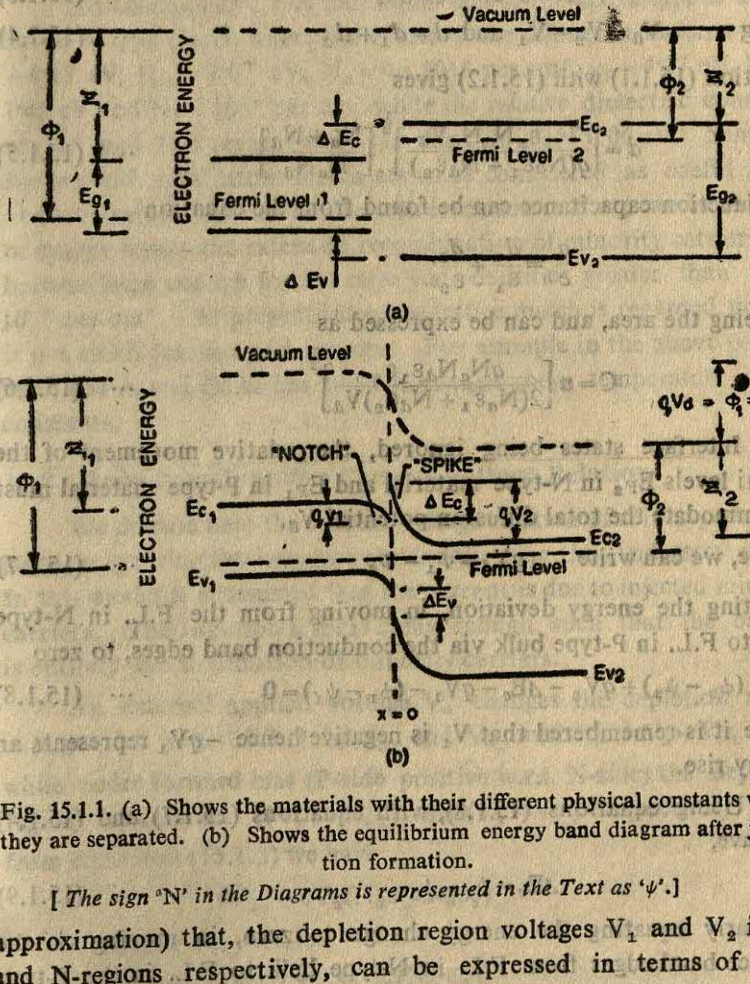


Fig. 15.1.1. (a) Shows the materials with their different physical constants when they are separated. (b) Shows the equilibrium energy band diagram after junction formation.

[*The sign* 'N' *in the Diagrams is represented in the Text as* 'ψ'.]

approximation) that, the depletion region voltages V_1 and V_2 in P- and N-regions respectively, can be expressed in terms of their supporting distances d_1 and d_2 as,

$$V_1 = -(qN_a d_1{}^2)/2\varepsilon_1 \text{ and } V_2 = (qN_a d_2{}^2)/2\varepsilon_2 \qquad \cdots \quad (15.1.1)$$

V_1 and V_2 being measured with respect to the interface as zero of potential.

Continuity of displacement perpendicular to the junction surface requires

$$N_a\, d_1 = N_a\, d_2 \qquad \cdots \quad (15.1.2)$$

Combining equations (15.1.1) and (15.1.2) gives the relation between the built in voltages

$$|V_1| N_a \varepsilon_1 = V_2 N_d \varepsilon_2 \qquad \cdots \quad (15.1.3)$$

Using $V_d = V_2 - V_1$ and $d = d_1 + d_2$ $\cdots$ (15.1.4)

equation (15.1.1) with (15.1.2) gives

$$d = \left[\frac{2\varepsilon_1\varepsilon_2 N_a N_d V_d}{q(N_a\varepsilon_1 + N_d\varepsilon_2)}\right]^{\frac{1}{2}} \left[\frac{N_a + N_d}{N_a N_d}\right] \qquad \cdots \quad (15.1.5)$$

The junction capacitance can be found from the equation

$$\frac{a}{c} = \frac{d_1}{\varepsilon_1} + \frac{d_2}{\varepsilon_2},$$

'*a*' being the area, and can be expressed as

$$C = a\left[\frac{q N_a N_d \varepsilon_1 \varepsilon_2}{2(N_a\varepsilon_1 + N_d\varepsilon_2)V_d}\right]^{\frac{1}{2}} \qquad \cdots \quad (15.1.6)$$

Interface states being ignored, the relative movement of the Fermi levels E_{F_2} in N-type material and E_{F_1} in P-type material must accommodate the total diffusion potential V_d.

Hence, we can write $qV_d = \Phi_1 - \Phi_2$ $\cdots$ (15.1.7)

Equating the energy deviation, in moving from the F.L. in N-type bulk to F.L. in P-type bulk via the conduction band edges, to zero

$$(\phi_2 - \psi_2) + qV_2 - \Delta E_c - qV_1 - (\phi_1 - \psi_1) = 0 \qquad \cdots \quad (15.1.8)$$

where it is remembered that V_1 is negative hence $-qV_1$ represents an energy rise.

Using equations (15.1.8), with equations (15.1.7) and (15.1.4) we have,

$$\Delta E_c = \psi_1 - \psi_2 \qquad \cdots \quad (15.1.9)$$

Similarly equating the energy change to zero, in moving via the valence band edges from F.L. in N-type bulk to F.L. in the P-type bulk, one has, using equations (15.1.7) with (15.1.4) and (15.1.9)

$$\Delta E_c + \Delta E_v = E_{g2} - E_{g1} \qquad \cdots \quad (15.1.10)$$

From equations (15.1.8) and (15.1.10) we have

$$\Delta E_v = E_{g2} - E_{g1} - (\psi_1 - \psi_2) \qquad \cdots \quad (15.1.11)$$

If interface states exist, the above idealised relation get modified. The energy bands at the interface moves on supplying the necessary interface charges but the ΔE_c step is still given by equation (15.1.9) though the equilibrium mobile population at the interface is largely controlled by these states.

An example of a P-N heterojunction of the above type is P-type Ge on N-type GaAs, where the values of the different constants are $\psi_1 = 4{\cdot}13$ eV, $E_{g1} = 0{\cdot}67$ eV, $N_a = 4 \times 10^{16}$ per cm^3, $\psi_2 = 4{\cdot}07$ eV, $E_{g2} = 1{\cdot}45$ eV and $N_d = 10^{16}$ per cm^3 while the relative dielectric constants are 16 and 11·5 respectively. In this case $\Delta E_c = 0{\cdot}06$ eV. Semiconductors with good lattice match are to be considered as useful pairs. At the interface between the two semiconductors there can be bending of energy bands and extensive recombination of minority carriers that become large enough for interface state densities greater than about 10^{12} per cm^3. At present interface lattice match is regarded good if it is a small fraction of 1 per cent. For example in the above pair Ge has 5·658 Å and GaAs has 5·654 Å as their room temperature lattice constants.

15.2 Anisotype Junction Current Voltage Relation

We discuss here the current voltage relation of anisotype heterojunction ignoring the interface states following the Anderson model. In this model it is assumed that the current is due to injected minority carriers. The injection level is small enough such that the current is entirely due to diffusion of minority carriers.

An external applied voltage V_a changes the depletion region voltages (cf. Fig. 15.2.1) by V_{a_1} and V_{a_2} such that $V_a = V_{a1} + V_{a2}$, while under forward bias (P-side positive w.r.t. N-side) the depletion region voltages are $(|V_1| - V_{a1})$ and $(V_2 - V_{a_2})$ respectively. Also from equations (15.1.3) we have

$$|\partial V_1| = \partial V_2 (N_d \varepsilon_2 / N_a \varepsilon_1)$$

$$\text{hence} \quad V_{a_2} = V_a/\theta \qquad \cdots \qquad (15.2.1)$$

$$\text{where} \quad \theta = [1 + (N_d\, \varepsilon_2 / N_a\, \varepsilon_1)]$$

That is, the fraction of the applied voltage effective on the wide gap side is more than that on the narrow gap side for $N_d\, \varepsilon_2 \angle N_a\, \varepsilon_1$. From Fig. 15.1.1(b) it can be visualised that, under forward bias condition, the barrier to electron flow from the wide gap material to the narrow gap material is much less than the barrier for hole flow in the opposite direction. Therefore, only the electron flow needs to be considered. Under equilibrium condition this current across the junction is zero. The barriers, on either side of the interface, presented to electron flow under equilibrium are qV_2 and $(\Delta E_c - q|V_1|)$ for flow from 2-1 and 1-2 mediums respectively. In

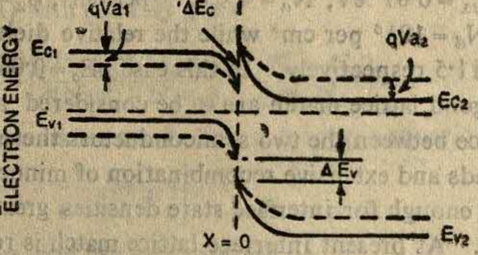


Fig. 15.2.1. Band diagram of **P-N** heterojunction. Bold lines show unbiased state, dotted lines show forward biased state.

equilibrium condition the two oppositely directed electron currents must equal, resulting in zero net current. This can be stated as

$$A_2 \exp(-qV_2/kT) = A_1 \exp-(\Delta E_c - q|V_1|)/kT \qquad \cdots \quad (15.2.2)$$

where A_1 and A_2 depend on the doping levels and effective masses of the electrons in the respective mediums. Considering the forward biased situation, the energy barriers would be $q(V_2 - V_{a2})$ and $[\Delta E_c - q\,(|V_1| - V_{a1})]$ respectively form the N- and P-type materials. Hence the unidirectional electron flux can be written as

$$A_2 \exp - q(V_2 - V_{a2})/kT - A_1 \exp - (\Delta E_c - q\,|\overline{V_1| - V_{a1}})/kT \qquad \cdots \quad (15.2.3)$$

which with equation (15.1.13) gives the forward bias current as

$$I = XA[\exp(-qV_2/kT)]\,[\exp(qV_{a2}/kT) - \exp(qV_{a1}/kT)] \qquad \cdots \quad (15.2.4)$$

where $A_2 = XA$ and the constant X is the fraction of the forward current that have enough energy to cross the barrier called the transmission coefficient, while A is the propagation parameter determined by diffusion by assumption and can be written as

$$A = aq\,\bar{n}_{po}\sqrt{D_n/\tau_n} \qquad \cdots \quad (15.2.5)$$

where D_n and τ_n are the diffusion constant and lifetime of minority carrier electrons. The ratio, of the two factors within the square bracket of equation (15.2.4), indicating the strenght of wide gap in relation to the narrow gap material in controlling the current, is $\exp(qV_a/kT)$. For V_a greater than a few times $\frac{kT}{q}$, the effect of V_{a2} is quite predominant. Hence the current is approximately exponential and can be written using (15.2.1) as

$$I = XA[\exp(-q V_s/kT)] [\exp(q V_a/\theta kT)] \qquad \cdots \qquad (15.2.6)$$

The experimental current in both forward and reverse bias is sometimes much higher⁽⁵⁾ or lower⁽⁶⁾ than given by the above equation. The transmission coefficient X in Anderson's model was, therefore, extensively, studied including its temperature dependence under both forward and reverse bias. These studies brought about a few models to describe the current flow mechanism through the interface region. These are :

(*a*) The emission model, which combines the classical kinetic model to determine the emission current with a diffusion model that determines the extent of minority carrier build up at the depletion edges. (*b*) The emission recombination model, that assumes a thin layer at the interface having a strongly disturbed lattice with very fast recombination characteristics, while holes and electrons reach the interface via thermal emission over their respective barriers. (*c*) In the tunnelling model, the current is obtained as the product of the tunnelling probability over the barrier and the incident electron flux of the wide gap N-region, while (*d*)—the tunnelling recombination model assumes band to band tunnelling coupled with recombination. Donnelly and Milnes⁽⁷⁾ has attempted a synthesis of these models, while Anderson⁽⁸⁾ has classified the forward and the reverse currents, each one into two types. The two types of forward current are thermal and non-thermal. Thermal type is one, in which the current carriers surmount the potential barrier to cross the junction and the current is of the form

$$I = A[\exp(-qV_{bo}/kT)][\exp(qV_a/\eta kT)] \qquad \cdots \qquad (15.2.7)$$

where V_{bo} is the barrier voltage at the thermal equilibrium, η is a barrier parameter determining the rate of flow, 'A' being a constant depending on the model. The non-thermal forward current is at least partly tunnelling in nature and the general expression for this current can be written as

$$I = B \exp[-\alpha(V_d - V_a)] \qquad \cdots \qquad (15.2.8)$$

where $\alpha = (4/3\hbar)(m_n^* \varepsilon_2/N_d)^{\frac{1}{2}}$ and V_d is the diffusion voltage.

Similarly, the reverse current is also classified as thermal and non-thermal. Thermal generation in or near the depletion region produces thermal current. Carriers generated within the depletion region are removed by electric fields while those outside diffuse to the depletion region thus either get recombined or add to the current flow. The non-thermal current under reverse bias is predominantly tunnelling current and is given by the expression

$$I = I_0 \exp - \beta (V_d + V_a)^{-\frac{1}{2}} \quad \ldots \quad (15.2.9)$$

where $$\beta = \frac{4}{3\hbar} \left(m_n^* \varepsilon_2 / N_d \right)^{\frac{1}{2}} E_{g2}^{\frac{3}{2}} \quad \ldots \quad (15.2.10)$$

the constants referring to the N-type materials.

The energy band diagrams of P-N heterojunctions are largely influenced by the parameters ψ, ϕ, E_g and charged interface states. Sharma Purohit⁽⁶⁾ has listed four distinct cases with altogether seven possibilities for abrupt anisotype junctions when interface states are ignored and has given their current voltage relations along with energy band diagrams. These four cases are for $E_{g1} < E_{g2}$, the subscript '1' referring to the P-type material.

(1) $\psi_1 < \psi_2$ with $\phi_1 < \phi_2$

Case (*a*) $\psi_2 > (\psi_1 + E_{g1})$ Case (*b*) $\psi_2 < (\psi_1 + E_{g1})$

(2) $\psi_1 < \psi_2 < (\psi_1 + E_{g1})$ with $\phi_1 > \phi_2$

(3) $\psi_1 > \psi_2$ with $\phi_1 > \phi_2$ and $(\psi_1 + E_{g1}) < (\psi_2 + E_{g2})$

Case (*a*) $V_{d1} > \Delta E_c$ Case (*b*) $V_{d1} < \Delta E_c$

(4) $\psi_1 > \psi_2$ with $\psi_1 < (\psi_2 + E_{g2}) < (\psi_1 + E_{g1})$

Case (*a*) $V_{d1} > \Delta E_c$ Case (*b*) $V_{d1} < \Delta E_c$.

The energy band diagrams of N-P heterojunctions can be drawn analogous to the P-N heterojunctions. Similarly there can be four distinct cases when the relative values of ψ, ϕ and E_g are considered. The above mentioned authors have listed these and have given their energy band diagrms alongwith current voltage relations. These four cases are for $E_{g1} < E_{g2}$, the subscript 'one' refering to N-type material.

(1) $\psi_1 > \psi_2$ with $\phi_1 > \phi_2$

Case (*a*) $\psi_1 > (\psi_2 + E_{g2})$ Case (*b*) $\psi_1 < (\psi_2 + E_{g2})$

(2) $\psi_1 > \psi_2$ with $\phi_1 < \phi_2$ and $(\psi_1 + E_{g1}) > (\psi_2 + E_{g2})$

(3) $\psi_1 > \psi_2$ with $\phi_1 < \phi_2$ and $(\psi_1 + E_{g1}) < (\psi_2 + E_{g2})$
Case (*a*) $V_{d1} > \triangle E_v$ Case (*b*) $V_{d1} < \triangle E_v$

(4) $\psi_1 < \psi_2$ with $(\psi_1 + E_{g1}) > \psi_2$
Case (a) $V_{d1} > \triangle E_v$ Case (*b*) $V_{d1} < \triangle E_v$

For hole flow on forward bias, from P-type to N-type materials, it can be easily shown that this can be written as

$$X_2 A_3 \, e^{-(qV_d + \Delta E_v)/kT} \; [\, e^{qVa/kT} - 1 \,]$$

where the constant A_3 refers to the mode of carrier propagation and X_2 stands for the fraction of the forward current carrying holes, that have enough energy to cross the barrier. Once within the N-type materials the flow can be taken as due to minority carrier diffusion, when the current due to this can be written for wide body as

$$I_p = X_2 \, a \, q \, \bar{p}_{no} \sqrt{\frac{D_p}{\tau_p}} \left[\exp - (qV_a + \triangle E_V)/kT \right] \left[\exp \left(\frac{qV_a}{kT} \right) - 1 \right] \quad \cdots \quad (15.2.11)$$

Using eqns. (15.2.6) and (15.2.11) the electron to hole current ratio can be given approximately as

$$\frac{I_n}{I_p} = \frac{X}{X_2} \frac{\bar{n}_{po}}{p_{no}} \sqrt{\frac{D_n \, \tau_p}{\tau_n \, D_p}} \, [\exp \{q(V_a - V_2) + \triangle E_V\}/kT] \quad \cdots \quad (15.2.12)$$

But except for the effective masses the ratio $(\bar{n}_{po}/\bar{p}_{no})$ can be written as

$$\frac{N_a}{N_d} \exp \{(E_{g2} - E_{g1})/kT)\}$$

Hence $(I_n/I_p) \simeq \exp (\triangle E_V + E_{g2} - E_{g1})/kT$. This shows that the injection efficiency of some heterojunctions can be very nearly unity.

15.3 Abrupt Isotype Junction

These are junctions of different band gap materials having same type of doping *i.e.* N-N or P-P junctions. A typical band diagram for N-N heterojunctions is shown in fig. (15.3.1). In this case the materials being of the same doping type, the wide band gap material shown is depleted and the narrow gap material is accumulated. Here, interface states are assumed to be absent. In order to find the relation between the diffusion voltages on the two sides of the metallurgical junction we assume here Boltzmann distribution over

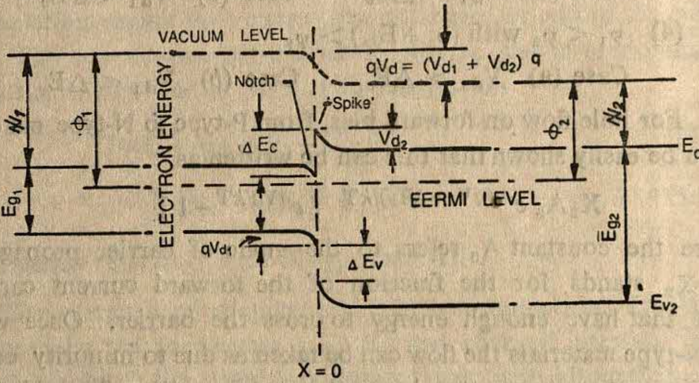


Fig. 15.3.1. Shows the equilibrium band diagram for an abrupt N-N heterojunction.

the available states. Far away from the interface the mobile carrier concentration that equals the dopant density can be written for the wide gap (subscript 2) material as :

$$N_{d2} = N_c \exp\left[-(E_{cf} - E_F)/kT\right]$$

where E_{cf} is the conduction band edge energy at this far away point. At a point lying in the depletion region of the wide gap material, there is a deficit of this number. This deficit $(N_{d2} - n)$ is responsible for the positive space charge. Hence the space charge equation can be written from $\left(1 - \frac{n}{N_{d2}}\right) = [1 - \exp(E_{cf} - E_c)/kT]$ as,

$$\frac{d^2V}{dx^2} = -\frac{N_{d2}\, q}{\varepsilon_2}\left[1 - \exp\frac{q}{kT}\left(V - V_{d2}\right)\right] \quad \cdots \quad (15.3.1)$$

Where V is the potential increasing to the right, V_{d2} the positive diffusion potential and ε_2 the permittivity of the wide gap region. Integrating equation (15.3.1) and evaluating the constant of integration by noting that, at $V = V_{d2}$, $\left(-\frac{dV}{dx}\right) = 0$ one has at the interface where $V = 0$

$$\left(\frac{dV}{dx}\right)^2 = \frac{2q\, N_{d2}}{\varepsilon_2}\left[V_{d2} + \frac{kT}{q}\left(\exp - q\, V_{d2}/kT - 1\right)\right] \quad \cdots \quad (15.3.2)$$

Similarly for the narrow band gap material of uniform doping N_{d1} the Poisson's equation is

$$\frac{d^2V}{dx^2}=\frac{qN_{d1}}{\varepsilon_2}\left[\exp\frac{qV}{kT}-1\right] \quad \cdots \quad (15.3.3)$$

where V is assumed to be zero far away from the interface within the narrow gap material. Equation (15.3.3) is integrated and the constant of integration can be evaluated by using, at $V=0$, $\left(-\frac{dV}{dx}\right)=0$, then one has at the interface,

$$\left(\frac{dV}{dx}\right)^2=\frac{2qN_{d1}}{\varepsilon_1}\left[\frac{kT}{q}\left(\exp qV_{d1}/kT-1\right)-V_{d1}\right] \quad \cdots \quad (15.3.4)$$

Using the above expressions for the interface fields in the two regions and equating for the continuity of displacements at the interface one can write

$$V_{d2}=\frac{N_{d1}\,\varepsilon_1}{N_{d2}\,\varepsilon_2}\left[\frac{kT}{q}\left(\exp\frac{qV_{d1}}{kT}-1\right)-V_{d1}\right]-\frac{kT}{q}\left[\exp\frac{-q\,V_{d2}}{kT}-1\right] \quad \cdots \quad (15.3.5)$$

Equation (15.3.5) shows that even for the same doping, so long V_{d1} is a few (kT/q), $V_{d2} \gg V_{d1}$ and the total diffusion potential $V_d=V_{d1}+V_{d2}$ is made up mostly of V_{d2}. Hence the transition region capacitance can be found using abrupt junction approximation and can be written as

$$C=a\left[\frac{q\,N_{d2}\varepsilon_2}{2V_d}\right]^{\frac{1}{2}} \quad \cdots \quad (15.3.6)$$

The above discussion shows that, of the total applied voltage V_a only a small part falls in the narrow gap material. Anderson has shown that the current voltage characteristics can be obtained by considering that the current flow is by thermionic emission process over a Schottky barrier. For this case the current is given by

$$I=B\exp\left(\frac{-q\,V_{d2}}{kT}\right)\left[\exp\left(\frac{q\,V_2}{kT}\right)-1\right] \quad \cdots \quad (15.3.7)$$

where B is an appropriate constant and V_2 is the voltage drop in the wide band gap material.

For large lattice mismatch and with one side heavy doping this model could not explain the experimental results. Kumar(9) has shown that for such cases diffusion model is better suited. This model predicts ohomic characteristics for certain conditions while rectifying for others but the former is rectifying always. However in the above model the interface states have not been considered. This calls for another model known as double Schottky diode model of Oldham and Milnes(10). In this the interface states are important and it is like two Schottky diodes connected back to back in series, each having its own saturation current. This requires double saturation which is not normally observed. A more general model proposed by Van Ruyven(11) considers the individual semiconductors to have their own surface states and divides the structure as due to two Schottky diodes formed of each semiconductor and its metal surface, having as if a metal contact between surface states. Current flow is due to tunnelling of carriers, when it is given by

$$I = I_s \exp [\alpha(V_d - V_a)] \quad \ldots \quad (15.3.8)$$

where I_s depends on the activation energy, on either side of the middle metal like surface between surface states and this metallic layer. A generalised model for isotype junctions is not yet proposed according to the knowledge of the author.

In isotype heterojunctions Sharma Purohit has divided the N-N junctions under four major groups and has given their current voltages relations.

These are,

Case 1. $\psi_1 > \psi_2$, $\phi_1 > \phi_2$ and $(\psi_1 + E_{g1}) < (\psi_2 + E_{g2})$

(*a*) $V_{d1} < \triangle E_c$ and (*b*) $V_{d1} > \triangle E_c$

Case 2. $\psi_1 > \psi_2$, $\phi_1 < \phi_2$ and $(\psi_1 + E_{g1}) < (\psi_2 + E_{g2})$

Case 3. $\psi_1 < \psi_2$, $\phi_1 < \phi_2$ and $(\psi_1 + E_{g1}) > \psi_2$

(*a*) $V_{d2} > \triangle E_c$ and (*b*) $V_{d2} < \triangle E_c$

Case 4. $\psi_1 > \varphi_2$, $\phi_1 > \phi_2$, $(\psi_1 + E_{g1}) > (\psi_2 + E_{g2})$

In above E_{g1} denotes the lower band gap.

Similarly the P-P heterojunctions are divided with E_{g1} used to signify the lower of the band gaps as :

Case 1. $\psi_1 > \psi_2$, $\phi_1 < \phi_2$, $(\psi_1 + E_{g1}) < (\psi_2 + E_{g2})$

Case 2. $\psi_1 > \psi_2$, $\phi_1 > \phi_2$, $(\psi_1 + E_{g1}) < (\psi_2 + E_{g2})$

Case 3. $\psi_1 > \psi_2$, $\phi_1 > \phi_2$, $(\psi_1 + E_{g1}) > (\psi_2 + E_{g2}) > \psi_1$

(*a*) $V_{d2} > \triangle E_v$ and (*b*) $V_{d2} < \triangle E_v$

Case 4. $\psi_1 < \psi_2$, $\phi_1 < \phi_2$, $(\psi_1 + E_{g1}) > \psi_2$

It would be an useful exercise to draw the energy band diagrams along the previously indicated lines.

15.4 Graded Heterostructures

There are many ternary and quarternary compound semiconductors in which the variation of chemical composition causes variation of band gap. For example the band gap variation with composition of some of the materials are depicted in fig. (15.4.1). If composition

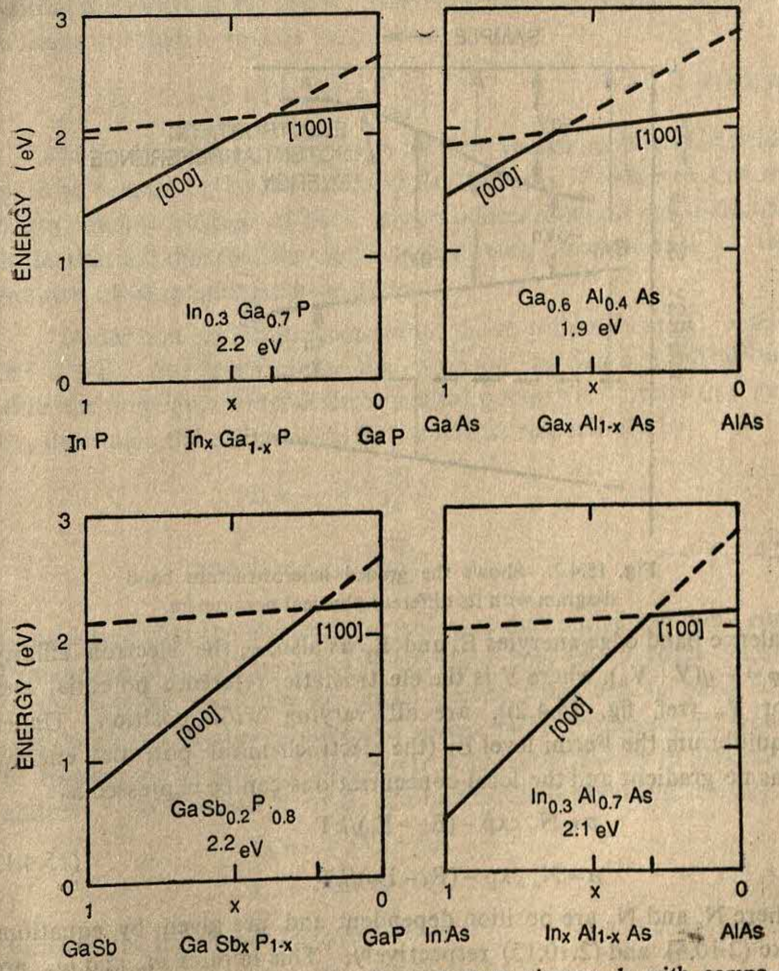


Fig. 15.4.1. The band gap variation of some ternary compounds with composition variation, direct to indirect transition composition is shown in the diagram.

variation along distance is gradual enough, the band structure is closely related to the band structures of the component alloys. In cases, when the relative variation of E_g is less than 5% in one hundred interactomic distances, the local macropotentials can be treated as if due to non-varying composition, which varies gradually with distance. Hence the energy band diagram representing the macroscopic potential variations due to composition variation can be represented as shown in fig. (15.4.2), where the conduction and valence band edge energies E_c and E_v as also, ψ, the electron affinity [$\psi = -q(V+V_n)$, where V is the electrostatic reference potential and for V_n (ref. fig. 15.4.2)], are all varying with position. Under equilibrium the Fermi level E_F (the electrochemical potential energy) has no gradient and the local concentrations can be expressed as

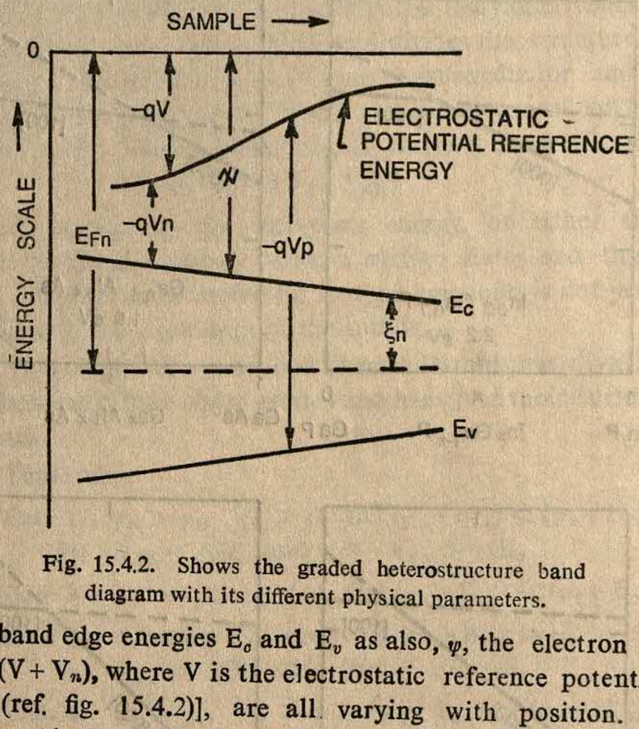


Fig. 15.4.2. Shows the graded heterostructure band diagram with its different physical parameters.

$$p = N_v \exp - (E_F - E_v)/kT$$

$$n = N_c \exp - (E_c - E_F)/kT \quad \cdots \quad (15.4.1)$$

where N_c and N_v are position dependent and are given by equations like (2.10.6) and (2.10.15) respectively. This implies m_n and m_p are position dependent. Hence the product

$$np = n_i^2 = N_c N_v \exp(-E_g/kT) \quad \cdots \quad (15.4.2)$$

varies as E_g varies with position x. Further, the equilibrium Fermi energy level E_i of intrinsic material of graded composttion is obtained by equating $n_i = p_i$ from equation (15.4.1) and is,

$$E_i = \frac{E_c + E_v}{2} + kT \log \frac{N_v}{N_c} \qquad \ldots \qquad (15.4.3)$$

But here E_i is position independent.

Hence taking the gradient of both sides of equation (15.4.3) and writing the values of N_c and N_v from equations (2.10.6) and (2.10.15) in terms of effective masses we can write

$$\frac{d}{dx}(E_c + E_v) = 3\,kT\left[\frac{d}{dx}\left(\log \frac{m_n}{m_p}\right)\right] \qquad \ldots \qquad (15.4.4)$$

This shows that band edge energy variations brings in fields given by equation (15.4.4). In 1957, H. Kroemer ⁽²⁾ showed theoretically that a change of band gap produces internal quasi-electric fields, that are different for electrons and holes, proportional to the gradiant of the respective band edges.

Under non-equilibrium condition, the so-called chemical potential $\xi_n = E_c - E_{Fn}$ determines the electron carrier concentrations while the non-equilibrium electrochemical potentials (imrefs) E_{Fn} and E_{Fp} determine the electron and hole currents respectively by

$$J_n = \mu_n\, n\, \frac{d\,E_{Fn}}{dx}$$

$$\text{and } J_p = \mu_p\, p\, \frac{d\,E_{Fp}}{dx} \qquad \ldots \qquad (15.4.5)$$

Considering that the electrons (or holes) at the conduction band edge (Valence band edge) experience electrostatic field due to space charge in the medium and there are quasi-electric fields due to band edge tilts (band gradings) even under no space charge condition, it would be appropriate to view the total electron potential energy in general as additive. So the local E_c or E_v can be expressed as (cf. fig. 15.4.2)

$$E_c = -q\,(V_n + V)$$

$$E_v = -q\,(V_p + V) \qquad \ldots \qquad (15.4.6)$$

Further, the imrefs can be expressed as

$$E_{Fn} = E_c + kT \log (n/N_c)$$

$$E_{Fp} = E_v - kT \log (p/N_v) \qquad \ldots \qquad (15.4.7)$$

In these two equations, the energies, potentials, concentrations and effective masses are all position dependent.

Now, writing for the quasi-electric fields

$$E_n = -\frac{\delta V_n}{\delta x} \text{ and } E_p = -\frac{\delta V_p}{\delta x} \quad \cdots \quad (15.4.8a)$$

and the electrostatic field $E = -\frac{\delta V}{\delta x}$, ... (15.4.8*b*)

the current expressions can be obtained from equations (15.4.5) with (15.4.6) and (15.4.7) as

$$J_n = n\mu_n\, q \left(E + E_n - \frac{3}{2}\frac{kT}{q}\frac{d}{dx}\log m_n\right) + D_n\, q\frac{dn}{dx} \quad \cdots \quad (15.4.9)$$

$$\text{and } J_p = p\mu_p\, q \left(E + E_p + \frac{3}{2}\frac{kT}{q}\frac{d}{dx}\log m_p\right) - D_p\, q\frac{dp}{dx} \quad \cdots \quad (15.4.10)$$

where D_n and D_p has been assumed to be field independent and given by $(\mu\, kT/q)$. As in uniform band gap semiconductor the current is composed of the diffusion current and drift current. But here the drift current depends on the fields associated with the electrostatic potential variation, the quasi-electric field associated with the conduction (Valance) band edge gradient and the carrier effective mass variation. The sum total fields are known as pseudo-fields.

The continuity equations can, as before, be written as

$$\frac{\delta p}{\delta t} = -\frac{1}{q}\frac{\delta J_p}{\delta x} - \frac{p - \bar{p}_n}{\tau_n}$$

$$\frac{\delta n}{\delta t} = \frac{1}{q}\frac{\delta J_n}{\delta x} - \frac{n - \bar{n}_p}{\tau_p} \quad \cdots \quad (15.4.11)$$

The interface states that may be arising from elemental boundaries in the bulk due to composition variation are small in number and can be ignored when the composition variation is much less than 5%, in about a hundred lattice constants. Strictly speaking, life time is also a function of position. In the next article we assume it to be position independent and examine the effect of band grading on the concept of the so-called diffusion length.

15.5 Minority Injection

Using a P-N junction of graded band gap materials minority carrier holes can be injected and its level can be maintained by forward biasing the junction. If the recombination is small then as before (cf. art. 4.3) the electron recombination replenishing current is not necessary. Under such circumstances the majority carrier electron current can be equated to zero to obtain the electrostatic field which can be written using equation (15.4.8) as

$$E = \frac{d}{dx}\left[V_n + \frac{kT}{q}\log\left(m_n^{\frac{3}{2}}/n\right)\right] \quad \cdots \quad (15.5.1)$$

Substituting this electrostatic field in the expression for minority hole current equation (15.4.10), one obtains the hole current using equation (15.4.7) as

$$J_p = -p\,D_p\,\frac{d}{dx}\left[\frac{E_g}{kT} + \log\frac{np}{(m_n m_p)^{\frac{3}{2}}}\right] \quad \cdots \quad (15.5.2)$$

where E_g is the space dependent band gap energy and $E_c - E_v = -q(V_n - V_p) = E_g = +qV_g$

Substituting, J_p from equation (15.5.2) in the hole continuity equation in (15.4.11) we have, on simplification, ignoring the equilibrium hole concentrations prior to injection, as

$$\frac{\delta^2}{\delta x^2} f(E_g, p, m) + \frac{d}{dx} f(E_g, p, m)\,\frac{d}{dx}\log p = \frac{1}{L_p^2} \quad \cdots \quad (15.5.3)$$

where $$f(E_g, p, m) = \left[\frac{E_g}{kT} + \log\left\{np/(m_n\,m_p)^{\frac{3}{2}}\right\}\right] \quad \cdots \quad (15.5.4)$$

It can be shown that an approximate solution for $f(E_g, p, m)$ of the form

$$f = \pm(c^2x - A\exp - x/c^2L_p^2)$$

is possible if L_p can be regarded as a constant, when, the hole concentration is given by

$$p = p_0 \exp(\pm x/c^2\,L_p^2)$$

This requires defining an effective diffusion length as $C^2L_p^2$; that is,

$$L_{eff} = L_p^2\left[\text{grad } E_g/2kT\left\{1 \pm \sqrt{1 + \frac{1}{L_p^2\,(\text{grad } E_g/2kT)^2}}\right\}\right] \quad \cdots \quad (15.5.5)$$

Equation (15.5.5) shows that whether grad E_g is positive or negative there are two possibilities of minority carrier distribution. For example, if grad E_g is positive two situations can happen as for minority carrier band edge tilt as shown here (for holes) in figs. (a) and (b). Minority injection at the end $x = 0$ in fig. 15.5.1 (*a*) will bring a normal diffusion opposed by valence band gradient while in fig. 15.5.1 (*b*) normal diffusion would be aided by the valcnce band gradient. In the former there will be a decrease while in the latter an increase of minority carriers with increasing x. If, say, in Ga_xAl_{1-x}. As the band

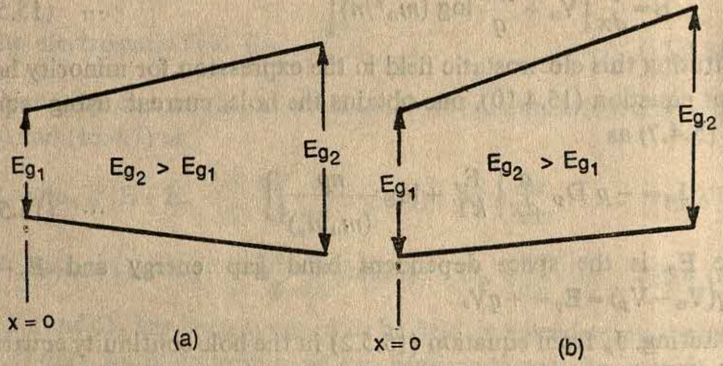


Fig. 15.5.1. Shows the two possibilities of the diffusion parameter.

gap changes due to composition variation from 1.45 eV to 1.75 eV in 0·3 microns and $L_p \sim 10^{-2}$ cm, then we have grad $\left(\frac{E_g}{2kT}\right) \simeq 2 \times 10^5/\text{cm} \gg 1/L_p \simeq 10^2/\text{cm}$. This shows the enormous dominance of band gradient over diffusion.

References

1. W. Shockley, U.S. Patent 2, 569, 347 (1951).
2. H. Kroemer, Proc. IRE 45, 1535, 1957.
3. R.L. Anderson, IBM. J. Res. Dev. 4, 283, 1960.
4. K.J. Adamsand Allen Nussbaum : A proposal for a new approach to Heterojunction theory. Solid State Electronics, Vol. 22, p 783—791, 1979.
5. A.G. Milnes and D.L. Fencht, Heterojunctions and metal semiconductor junctions, Academic Press 1972, p. 40.
6. B.L. Sharma and R.K. Purohit, Semiconductor Heterojunctions, Pergamon Press 1974, p- 4.
7. J.P. Donnelly and A.G. Milnes, Proc. IEE 113, 1468 (1966).

8. R.L. Anderson, Proc. Int. Conf. on Phys, Chem of Semicond. Heterojunctions, Vol. II, p. 55, Akademiai Kiato, Budapest, 1971.

9. R.C. Kumar, Int. J. Electronics 25, 239 (1968).

10. W.G. Oldham and A.G. Milnes, Solid State Electronics 7, 153 (1964).

11. L.J. Van Ruyven et al, Solid State Electronics 8, 631 (1965).

Problems

1. Draw the energy band diagram for a P-type Si forming a junction with N-type Ga P. Use $\psi_1 = 4{\cdot}0$ *eV*, $\psi_2 = 4{\cdot}3$ *eV* $E_{g1} = 1{\cdot}1$ *eV*, $E_{g2} = 2{\cdot}25$ *eV*, $\phi_2 = 4{\cdot}6$. Find the conduction band and valence band energy steps. Group this according to the division mentioned in the text.

   Ans. $E_c = 0{\cdot}3$ *eV* $E_v = 1{\cdot}45$ *eV*. Case (2) P-N.

2. Draw the energy band diagram for P-type AlSb forming a junction with N-GaSb. Use $\psi_1 = 4{\cdot}10$*eV*, $\psi_2 = 3{\cdot}7$ *eV*, $\phi_1 = 4{\cdot}3$ *eV*, $\phi_2 = 5{\cdot}1$ *eV*, $E_{g1} = 0{\cdot}7$ *eV* and for AlSb $E_{g2} = 1{\cdot}6$ *eV*.

Chapter 16

PHOTO DETECTORS

In this chapter we study the optical signal detection devices. The information carrier is EM radiation. The photon energy $E = hf$ associated with the EM rediation excites charge carriers in detectors. The movement of those carriers inside the detector induces a current in the external circuit. We normally refer to the detection wavelength. The optical frequency ranges of interest, their corresponding wavelengths and associated photon energies are shown in Table (16.1).

Table 16.1

Range	Wavelength in microns	Frequency in Hz	Photon energy in electron volts
Infra-red	200—0·7	$1{\cdot}5 \times 10^{12}$—$0{\cdot}43 \times 10^{15}$	$6{\cdot}2 \times 10^{-3}$—1·77
Visible	0·7—0·4	$(0{\cdot}43$—$0{\cdot}75) \times 10^{15}$	1·77—3·1
Ultra-violet	0·4—0·002 (20Å)	$0{\cdot}75 \times 10^{15}$—$1{\cdot}5 \times 10^{17}$	3·1—620.

There are two basic principles of detection. These are the video detection and heterodyne detection while photodetectors respond to the intensity of radiation incident on it. The optical signal is normally on-off type or amplitude modulated. In video detection, the detector gives a signal frequency component of current the peak value of which in proportional to twice the product of carrier power and modulation index. In heterodyne detection, on the other hand, the received optical signal is added with a strong local signal, of a frequency, higher than the received signal frequency by the intermediate frequency (IF). This then being incident on the detector gives a current at the IF that is approximately proportional to the product of the square root of the local oscillator signal power and the received signal power ($\sqrt{P_s P_L}$), which contains the information. Heterodyne detection is most useful for the infra-red region because of the availability of strong monochromatic local oscillator lasers.

The photodetector response can be described by the responsivity (or sensitivity). This depends on the wavelength response as also on the speed of response of the detector. It is defined by the ratio of rms value of current or voltage at the detector output to the rms value of monochromatic optical power incident on the detector. In the detector one needs to know the minimum detectable power of the signal in presence of noise. The detector frequency response and transit time delay are also important. A figure of merit called the noise equivalent power (NEP) is used frequently. This is, the rms value of the sinusoidally modulated optical radiation incident on the photodetector that gives rise to an rms signal equal to the noise voltage output of the detector. Another important figure of merit used for infra-red detector is called the detectivity. This is defined as $D^* = (A \triangle f)^{1/2}/(\text{NEP})$ where A is the photodetector useful area and $\triangle f$ the associated bandwidth. To avoid ambiguity symbols like $D^* (\lambda, f, 1)$ or $D^* (T, f, 1)$ are used, to state the carrier source λ, or reference temperature T in degree Kelvin (if the source is characterised as black body). The modulation frequency is f, the reference bandwidth for performance comparison being taken as 1 Hz.

There are four types of important detectors. These are : (1) The photo-emitter type, (2) the photoconductor, (3) the photodiode and (4) the avalanche photodiode (APD) detectors. Recently the heterojunctions with different techniques and constructional features have made further improvement in detector performance in sensitivity and wavelength range.

16.1 Photoemissive Detectors

These are of three types, the Vacuum photodiodes, the photomultipliers and the gas-filled photo-tubes. In all these devices light detection is achieved by photoexcitation of electrons over the surface barrier of a photo cathode. In gas-filled tubes gas at about 0·1 torr is enclosed in a chamber containing the photoemissive cathode. Light admitted through glass or quartz window excites electrons from the cathode that are led to the anode at which a large positive voltage is applied. The moving electrons ionise the gas inside. The current flows through a load resistor that gives the detected signal. The slowly moving ions restrict the frequency range of these detectors below 10 KHz.

Photomultipliers also use photoemissive material. Since their invention photomultipliers have been developed to great perfection. These are usable for ultra-violet, visible and near infra-red regions of the spectrum. It consists of a photocathode (C) coated with photoemitting material and a number of electrodes called dynodes, which are lebelled 1 through 10 as in fig. (16.1.1) The dynode potentials are made progressively higher with respect to the cathode. The potential difference between consecutive dynodes being about 100 volts. The last electrode (A) is called the anode. The whole electrode system is placed inside a vacuum envelope. The optical ux incident on the photocathode causes electron emission from it.

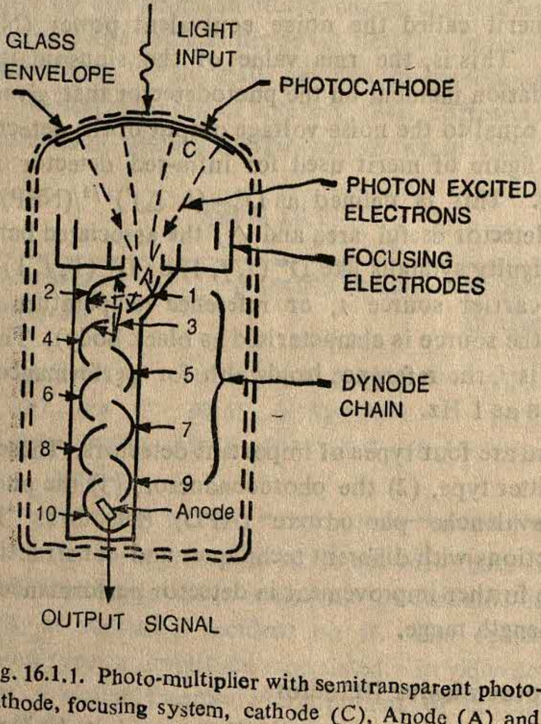


Fig. 16.1.1. Photo-multiplier with semitransparent photocathode, focusing system, cathode (C), Anode (A) and dynodes as numbered.

The electrons are subjected to electrostatic focussing by the focussing enclosure as shown in the figure. The electrons are accelerated due to the potential difference between the electrodes. The accelerated electrons incident on dynodes cause secondary electron emission. This causes multiplication of electrons at each dynode. The repeated multiplication causes amplification of the initial cathode emitted current by a large factor. Finally these electrons are collected by the anode. Assuming, the average secondary emission phenomena producing a multiplication by M at each dynode and that there are n dynodes, the cathode emitted primary electrons get multiplied by a factor $G = M^n$, for typical values $n = 8$ to 10 and $M = 4$ to 5 ; $G \sim 10^6$.

The are two types of photocathodes : (1) the semitransparent one, which is a thin film of photoemissive material deposited on a transparent substrate and (2) the reflection type, that uses thick films of photoemissive material on conducting substrates. The time needed for emission is sufficiently small so that the photoemission current can follow the modulation frequencies even far into the microwave region. An optical power P corresponds to (P/hf) quantas incident per second. If, on the average, each incident quanta produces η charge carriers, the photo-emitted rms current in video detection for modulation index m is*

$$I_{ph} = \sqrt{2}m\,\eta\,q\,P/hf \qquad \cdots \qquad (16.1.1)$$

The factor η is called the quantum efficiency. Most conventional photo-emitters are covered with a monolayer of cesium which lowers the electron affinity, that in turn leads to higher quantum efficiency. This extends the cathode photo-response towards longer wavelengths.

Another important parameter of the photo-emitters is the dark current i_a that arises from random thermal excitation of electrons, excitation by cosmic ray or by unchecked radio-active bombardment.

---

* In video detection $i(t) \propto [E_c(1 + m\cos\omega_m t)]^2$ while the average photocurrent is proportional to the average power P, hence $\bar{i} = \frac{Pq\eta}{hf}$, that occurs at $m = 0$ also. Hence the constant of proportionality can be eliminated and the signal current with gain G can be written as $i(s) = G.2\,m\bar{i}\cos\omega_m t$ and is

$$i_s = \frac{GPq\eta}{hf}(2m\cos\omega_m t)$$

Photocathodes for different wavelength ranges are obtainable with emitter compositions like GaAs—Cs_2O between 0·4 micron to 0·9 micron with $\eta \sim 0{\cdot}15$, In $As_{0\cdot15}$ $P_{0\cdot85}$—Cs_2O with $\eta \sim 0{\cdot}02$ at 1·06 microns and 0·15 at 0.7 micorns (cf. ref. 1). The dark current in the former is of the order of 10^{-16} amp per sq. cm. and in the latter 10^{-11} amp per sq. cm. at 23°C. The dark current is given by Richardson equation $I = AT^2 \exp(-\phi/kT)$ where ϕ is the work function of the coating material. The long wavelength frequency limit is determined by the cathode material work function. It is apparent that the emitting material also determines the response sensitivity.

The noise contributions in photo-multipliers are :

1. The cathode shot noise that arises from the random emission process. The mean square value of the shot noise current can be written in terms of the emitted mean current I arising out of incident mean power as*

$$\overline{i^2_{n1}} = 2G^2 q\,(\bar{I} + i_d).\ \Delta f \qquad \cdots \qquad (16.1.2)$$

where i_d is the dark current.

2. The Johnson noise that arises from the load conductance G_L of the photomultiplier which includes the amplifier input following the photomultiplier. The mean square noise current due to this cause $\overline{i^2_{n2}} = 4kT\,G_L\,\Delta f$ $\cdots$ (16.1.3)

3. The dynode shot noise that arises due to randomness of the secondary emission process. It does not experience full gain of the photomultiplier, hence is small and can be ignored. Considering the above, the signal to noise power ratio of the photomultiplier detector can be written as

$$(S/N) = (G^2\,I_{ph}^{\,2})/(\overline{i_{n1}^{\,2}} + \overline{i_{n2}^{\,2}}) \qquad \cdots \qquad (16.1.4)$$

In the case of video detection, if the signal power is low enough even for $m=1$, the noise due to dark current predominates

---

* If $F(\omega)$ is the Fourier transform of each random event ; the average power per unit bandwidth is given by $2N\ |F(\omega)|^2$, where N is the average number of events per second. This is known as Carson's theorem. This gives the mean square current due to shot noise as

$$i^2 n = 2qI\ \Delta f.$$

and the minimum detectable power (provided the Johnson noise is kept low enough) is obtained by equating the signal to noise ratio to unity as

$$P_{min} = (hf\sqrt{i_d\,\Delta f}/\eta\sqrt{q}) \qquad \cdots \qquad (16.1.5)$$

In heterodyne detection the major cathode shot noise arises from the local oscillator power since the *d.c.* cathode current (cf. Art. 16.2) due to local oscillator signal field is large. Hence the output shot noise mean square current is

$$\overline{i_{n1}^{\,2}} = 2\,G^2 q[i_d + (P_L\,q\eta/hf)]\,\Delta f \qquad \cdots \qquad (16.1.6)$$

where P_L is the local oscillator power. As the local oscillator power is increased, all other noise contributions are dwarfed by the local oscillator power contributed shot noise. Under this condition, since $S/N = [2G^2\,P_S P_L\,(q\eta/hf)^2]\,/(\overline{i_{n_1}^{\,2}} + \overline{i_{n_2}^{\,2}})$ one has,

$$(S/N) = P_s\,\eta/hf\,\Delta f \qquad \cdots \qquad (16.1.7)$$

where P_S is the incoming signal power.

The minimum detectable power is obtained by equating the signal to noise ratio to unity as

$$P_{min} = (hf\Delta f/\eta) \qquad \cdots \qquad (16.1.8)$$

This is called the 'quantum limit' of detection and is the power that is equivalent to one photon per unit bandwidth *i.e.* within one unit of system resolution time. For $f = 1$ Hz this is of the order of 10^{-18} watt. This limit is achievable in heterodyne mode of detection using photomultipliers. The speed of response is determined by the transit time dispersion of electrons between dynodes, the capacitance and output circuit transit time. The bandwidth becomes proportional to the dynode voltage, inversely proportional to their spacing and square root of their number. Improved dynode structures with bandwidths in the GHz range are also obtainable.

In vacuum photodiode, electrons emitted from a photo cathode into vacuum are collected by anode to produce a signal current through a load resistor. These are normally used for intense laser beams. These have been used upto 10 GHz by mounting in MW circuits with (transit time minimising) rather high anode voltages. Specially designed co-axial photodiodes can detect over the frequency range from 1·5—51 GHz. For weak signals sensitivity limitation is overcome by coupling the electron beam to a high Q cavity of a photo-klystron, when it becomes useful for narrow bandwidths. Shophisticated helical slow wave structures have been used in the frequency band from 10—20 GHz range.

16.2 Photoconductive Detector

A photoconductor detector is a piece of semiconductor bulk or a thin film between two ohmic contacts. When it is irradiated with light, there may be band to band transition producing hole electron pairs or there can be transition between bands and energy levels in the forbidden gap, that may have been incorporated for this purpose. In either case the sample conductivity increases, since the number of mobile carriers increases due to illumination. The former type is called the intrinsic photoconductor while the latter is extrinsic. A bias voltage transforms light induced conductivity modulation into current modulation in a load. The long wavelength limit is determined by the minimum energy gap through which carrier transition occurs and is given by $hf_c = qV_g$, where f_c is the cut-off frequency. The wavelength in microns is given by

$$\lambda_c = (1{\cdot}242/V_g) \quad \cdots \quad (16.2.1)$$

where V_g is the energy gap voltage in volts. Photons having higher energies are absorbed. In extrinsic photoconductors, impurities either deep or shallow, are incorporated within the band gap. Here the transition between the impurity level and band edges becomes useful. For longer wavelengths lower ionisation energy impurities are used, such as Cu in Ge. These act as acceptor with ionisation energy $E_i = 0{\cdot}04$ ev. This corresponds to a cut-off $\lambda_c = 32$ microns. At room temperature range the free particle energy being $(3KT/2) = 36$ mev, the photoconductor is cooled to avoid thermal excitation. Similarly In is a deep impurity in Si and at 77°K or less and is good for the 3 to 5 micron range detection. Deep impurities like Au, Zn, S, In and the shallow ones like B, Al, As, Sb have been largely used to cover the range from a few microns to about 30 microns.

The performance of the intrinsic photoconductor depends on the material through its absorption coefficients at different wavelengths. The absorption co-efficients α (λ), for some common materials are presented in fig. (16.2.1). The absorption implies that the fractional decrease of photon flux N for a penetration of an elemental depth can be expressed as

$$\frac{dN}{N} = -\alpha\, dx \qquad \cdots \quad (16.2.2)$$

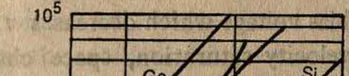


Fig. 16.2.1. Absorption co-efficient *vrs.* λ in microns for some common semi-conductor.

This implies that the photon flux decreases following an exponential law such as

$$N = N_0 \exp(-\alpha x) \qquad \cdots \quad (16.2.3)$$

For high quantum efficiency, the photoconductor thickness in the direction of incident light should be at least $(2{\cdot}3/\alpha)$ for 90% of the light to get absorbed.

Though the generation of both holes and electrons occur due to optical excitation, one type of carrier may be trapped or made ineffective in impurity or defect states as in CdS, CdSe or as it occurs in extrinsic semiconductors. For a semiconductor in which one carrier conductivity change is dominant, a photon generated η number of carriers appears as ηT_0, before it recombines since the generation rate $g = \hat{p}/T_0$, T_0 being the carrier life time. The number of times these carriers flow per second is $(1/t_1)$, where the transit time $t_r = L/\mu E = L^2/\mu V$, L being the semiconductor length and V the applied voltage. Hence the conductivity change appears to have increased by the factor T_0/t_r. This is precisely the number of times the carriers make passes. This gain is referred to as quantum gain or current gain, which can be as high as 10^5. The current gain of the photoconductor increases with bias voltage which decreases t_r, until any of the events like carrier velocity saturation, space charge limited current flow, exhaustion of the ohmic contact source or breakdown occurs.

In intrinsic photoconductors at low frequencies the light induced current is given by

$$I = \frac{qP}{hf}\left(\eta_n \frac{T_n}{t_{rn}} + \eta_p \frac{T_p}{t_{rp}}\right) \qquad \dots \quad (16.2.4)$$

where P is the total optical power absorbed in the photoconductor, t_{rn} and t_{rp}, η_n and η_p are the transit times and quantum efficiencies of the electrons and holes respectively. In the intrinsic photoconductors both electrons and holes remain mobile and contribute to conductivity. The sample having low initial conductivity and space charge neutrality requiring the excess concentrations of both type of carriers at each point to follow each other, the life time becomes same for both types of carriers. The current expression (16.2.4) may lead one to expect higher sensitivity with intrinsic photodetectors. But as long as neutrality is maintained at higher light levels (since generation recombination noise is appreciable), both types of carriers flow with a common mobility (ambipolar mobility) towards the same contact in pairs, thus limiting the sensitivity. On the other hand for unequal hole and electron densities *i.e.* in doped semiconductors and for low light intensities) the ambipolar mobility reduces to the minority carrier mobility. Thus providing an one-carrier dominance in photoconductivity and the current gain saturates. Application of high voltage sweeps out the minority carriers with increased current gain and high speed response is obtained. To combat the decrease in current gain trapping centres are introduced that immobilise one type of carrier or one contact is made blocking inhibiting the flow of one carrier type as the holes in the emitter of an N^+-N-P-N transistor. Thus essentially high current gain is obtained with only one type of mobile carriers.

In heterodyne detection the photogenerated signal current can be written* as

$$I_p(\omega) - \frac{A\,\eta\,q\,P_1}{hf\sqrt{2}}\,\frac{T_0}{t_r\sqrt{(1+\omega^2 T_0{}^2)}} \qquad \cdots \quad (16.2.5)$$

where A is the photoconductor area and $P_1 = 2\sqrt{P_s P_L}$, P_s and P_L being the received signal and local oscillator signal powers incident per unit area. Hence the available power at a matched load (cf. fig. 16.2.2) at high frequencies ($\omega T_0 \gg 1$) is

$$P_a = \left(\frac{P_1 A \eta q}{hf.\omega t_r}\right)^2 \Big/ 8G_L \qquad \cdots \quad (16.2.6)$$

As for the noise mechanism in a photoconductor, the carrier generation is a hit and mis process, in addition ; the recombination characterising life time T_0 is a statistical quantity**. These considerations give a mean square generation recombination current in a $\triangle f$ band width*** as

$$\overline{i^2{}_{gr}} = S(\omega)\triangle f = \frac{4q\bar{I}T_0\triangle f}{t_r(1+\omega^2 T_0{}^2)} \qquad \cdots \quad (16.2.7)$$

For the Johnson noise on the other hand the mean square current is

$$\overline{i^2{}_{n2}} = 4KT_e G_L \triangle f \qquad \cdots \quad (16.2.8)$$

---

*The local oscillator and optical signal field summed is, $V(t) = \mathrm{Re}\,[E_s \exp j\,\omega_s t + E_L \exp j\,(\omega_s + \omega)\,t]$, which on detection gives $i\,(t) = C\;V\,(t)\;V^*(t) = C(E_L{}^2 + E_s{}^2 + 2\,E_L E_s \cos \omega t)$. The mean d.c. current is obtained from $C\,(E_L{}^2 + E_s{}^2)$, which can be written, for area A assuming P stands for power per unit area, as $A[(P_s + P_L)\,\eta q\;T_o]/(t_r\,hf)$,

Hence $i(t) = A\,\frac{\eta q\,T_0}{hf\,T_r}\left[P_s + P_L + \frac{2\sqrt{P_s\,P_L}}{\sqrt{1+\omega^2 T_0{}^2}}\cos \omega t\right]$, where frequency dependance of current due to life time is taken into account.

**The life time is a statistical quantity, in that, it is a amean value of a distribution of events in time, while the total probability of its existence is unity. Hence this probability can be expressed as $g\,(\tau) = [\exp\,(-t/T)]/T$ such that

$$\int_0^\infty g(\tau)dt = 1.$$

***During the lifetime $i(t) = q\bar{v}/L$ while on either side of this time domain $i(t) = 0$. This has a fourier transform $I\,(\omega,\;T_o)$. Hence its mean square value is $\overline{I_s{}^2(\omega,\,T_o)} = \int_0^\infty \left| I\,(\omega,\;T)\right|^2 g(T)\;dT = 2\;(qT_0/t_r)^2/(1+\omega^2 T_0{}^2)$. The spectral density function from Carson's theorem $S\,(\omega) = 2\bar{N}\left|\overline{I_s\,(\omega,\,T_o)}\right|^2$, where $\bar{N}$ is the average number of carriers generated per second per unit volume. But the average current is $I = A\;\;\bar{N}\,q\,T_0/t_r = A\,q\,\eta\;T_o\;\;(P_s + P_L)/(hf\,t_r)$. hence the result.

where T_e is the equivalent noise temperature of the diode conductance G_1 and load conductance G_L. At low frequencies there is current dependent ($1/f$) noise due to surface and contact imperfections which is ignored in the present discussion. Fig. (16.2.2) shows

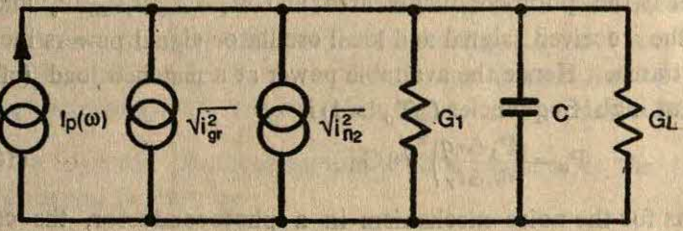


Fig. 16.2.2. Equivalent circuit of a photo-conductive detector.

the diode equivalent circuit with matched load and the noise and signal current generators. The capacitance C stands for the total terminal capacitor including the amplifier input. From equations (16.2.5), (16.2.7) and (16.2.8) the signal to noise power ratio for photoconductive detector can be written using

$P_0 = (P_s + P_L)$ and $P_1 = mP_0$ as

$$S/N = \frac{A\eta m^2 P_0}{8hf\Delta f} \Big/ \left[1 + \frac{KT_e}{q}\frac{t_r}{T_0}(1+\omega^2 T_0^2)\, G_L/I\right] \quad \cdots \quad (16.2.9)$$

When the generation recombination noise is large $(S/N) \simeq (P_s \eta A/2hf \times \Delta f)$. The minimum detectable signal under this condition is obtained by equating $(S/N) = 1$ as $P_{s\,\min} = (2hf\,\Delta f/\eta)$. This is twice the photomultiplier heterodyne detection limit, but η in semiconductors is of the order of unity while in photomultiplier it is $\sim 0{\cdot}3$. Measurement of NEP shows that high sensitivity approaching the quantum noise limit can be reached over wide IF bandwidths with Ge : Cu(Sb), photoconductive detectors at 4·2°K (c.f. ref. 1). Fig. (16.2.2) shows that the detector bandwidth is determined ultimately by the time constant (Req C), where $Req = (1/G_1 + G_L)$, since transit time limitation can be avoided by decreasing the length L and increasing bias voltage V. By calculating the load current it can be easily seen that the current gain bandwidth product is given by $(1/2\pi t_r)$.

16.3 Photodiodes

Solid state photodiodes essentially consist of a reverse biased P-N junction, a P-I-N, a metal semiconductor junction or a heterojunction diode. In the depletion region of these diodes the electric field is high. In these diodes absorbed photons excite electron-hole pairs through band to band transition. Electron-hole pairs are generated in or near the depletion region. The minority carrier flow is such that, electrons flow to the N-side while the holes to the P-side (cf. fig. 16.3.1). The depletion region field assists their flow. The minority carriers excited within a few diffusion lengths within the bulk, on either side of the depletion region, gets diffused to the junction edge. These are then assisted by the drift field and get collected on the respective sides where they are majority carriers. This gives the photo current in junction diodes. The d.c. current voltage characteristic can be given by the equation

$$I = I_L - I_0[\exp(-qv/kT) - 1] \qquad \cdots \qquad (16.3.1)$$

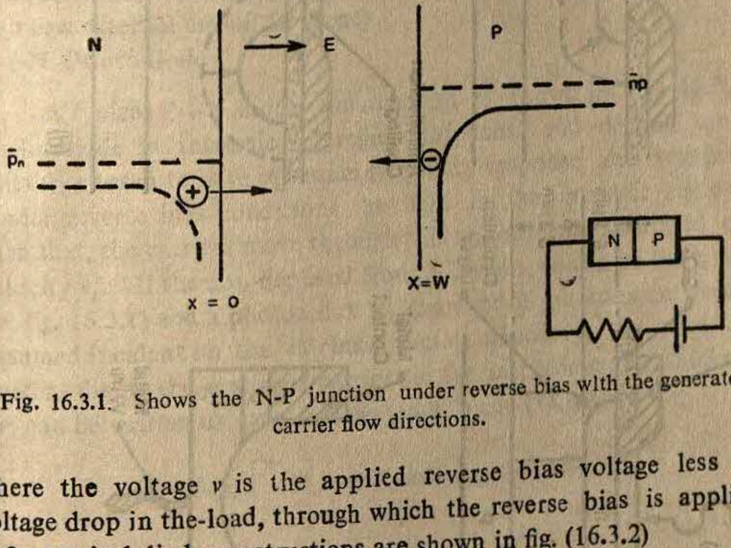


Fig. 16.3.1. Shows the N-P junction under reverse bias with the generated carrier flow directions.

where the voltage v is the applied reverse bias voltage less the voltage drop in the-load, through which the reverse bias is applied. A few typical diode constructions are shown in fig. (16.3.2)

As in photoconductor, the wevelength response is determined by the absorption coefficient as a function of λ of the diode material and the carrier collection. The carrier collection is determined by the depth of the P-N junction, minority carrier diffusion length and surface recombination velocity. The long wavelength cut-off is determined by the forbidden gap of the semiconductor. In short wavelength range the absorption coefficient being high, the high energy photons are mostly absorbed within the semiconductor very

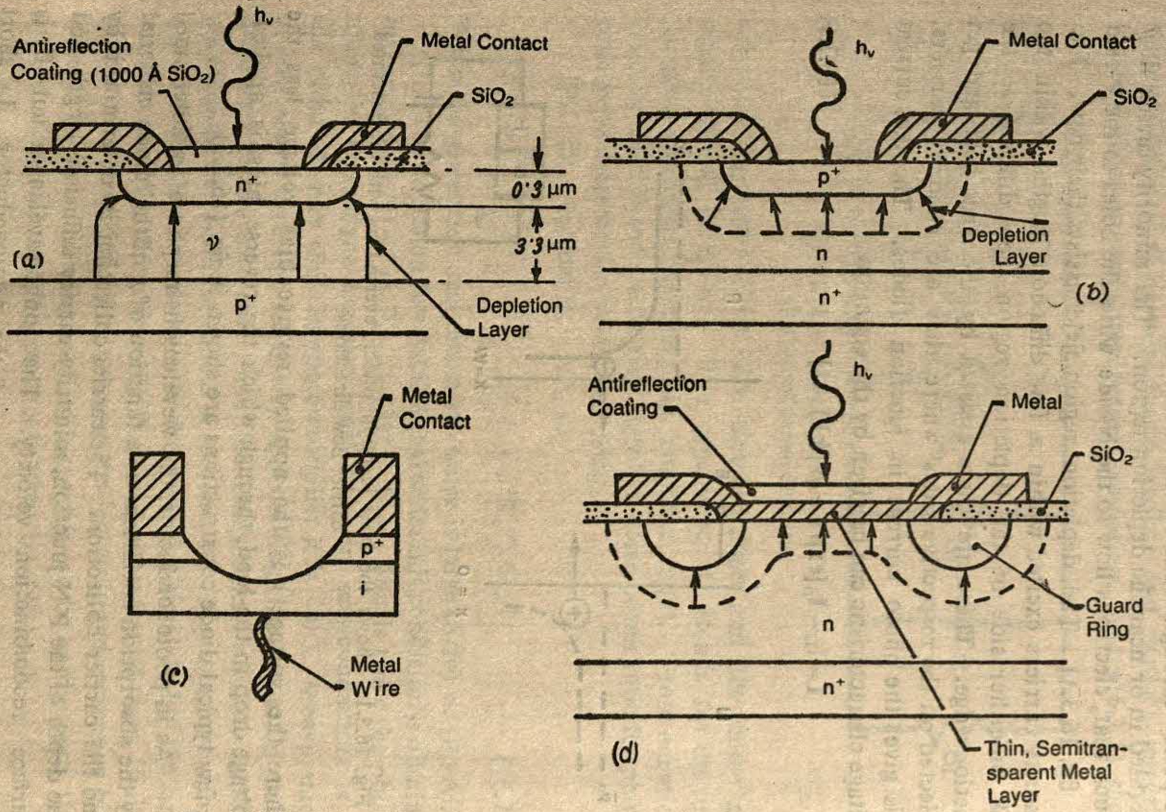


Fig. 16.3.2(a—d). Construction of some photodiodes described in the text.

near the surface of incidence of the light. So the short wavelength cut-off is influenced by the collection of the minority carriers (generated just within the surface) at the remote electrodes. This means that the minority carrier surface recombination velocity needs to be small. Additionally the collection can be improved by the application of large reverse bias (less than junction breakdown voltage) that gives large depletion region extending to very near the surface. This makes faster transit of carriers through the depletion region, due to increase of depletion region field, resulting in lesser recombinations in the collection process. If the depletion region is too wide, this may impair the frequency performance due to increased transit time through this region. The depletion region should not be too thin either, because in that case the depletion capacitance will give a large passive RC time constant of the diode, resulting in a low external cut-off frequency.

P-I-N Photodiode :

A typical P-I-N diode construction is shown in fig. (16.3.2a). In this diode in intrinsic (ν) region thickness and doping can be suitably chosen to have optimum frequency response and sensitivity. Under reverse bias conditions the field in the intrinsic region is such that, the carriers move through this region with saturated drift velocity v_s. If the non-depleted front N-region thickness is ignored (cf. fig. 16.3.1) and a photon flux N_0 (number of photons/cm²/sec) is assumed incident on the intrinsic region boundary, the total drift current due to the generated carriers in the intrinsic region thickness, W, can be written using equations (16.2.2) and (16.2.3) as

$$I_{dr} = Aq\eta \int_0^w N_0 \, \alpha e^{-\alpha x} \, dx = Aq\eta \, N_0(1 - e^{-\alpha w}) \quad \ldots \quad (16.3.2)$$

where η is the quantum efficiency *i.e.*, the number of generated carriers per photon. The minority carriers generated in the P-bulk of the diode diffuse through the bulk, to reach *i*-region boundary. The bulk thickness is more than a few diffusion length which acts as the substrate. The thickness of the substrate being large, the carrier distribution is exponential. The minority carrier electron concentration at $x = W$ is zero, since the diode is reverse biased and at $x = \infty$, it is $\bar{n}_p$, the thermal equilibrium value. Hence the minority carrier concentration in the P bulk, beyond the depletion boundary ($x > \mathrm{W}$) can be written as

$$n = \bar{n}_p - [\bar{n}_p + \mathrm{C} \exp(-\alpha \mathrm{W})] \exp(\mathrm{W} - x)/\mathrm{L}_n + \mathrm{C} \exp(-\alpha x) \quad \cdots \quad (16.3.3)$$

where C is a constant required to satisfy the continuity equation for minority carriers, with diffusion, generation and recombination only, which can be written with $\hat{n} = (n - \bar{n}_p)$ as

$$\mathrm{D}_n \frac{\delta^2 \hat{n}}{\delta x^2} - \frac{\hat{n}}{\gamma} + \eta \, \mathrm{N}_0 \, \alpha \, e^{-\alpha x} = 0 \quad \cdots \quad (16.3.4)$$

The constant C will be found to be

$$\mathrm{C} = \frac{\eta \mathrm{N}_0 \ \alpha \ \mathrm{L}_n^2)}{\mathrm{D}_n(1 - \alpha^2 \mathrm{L}_n^2} \quad \cdots \quad (16.3.5)$$

Ignoring the thermal generation and recombination, which for the wide body is $(q \, \bar{n}_p \, \mathrm{D}_n)/\mathrm{L}_n$, the diffused optically generated current reaching the depletion boundary at W, obtained from $\mathrm{AD}_n \, q \left(\frac{dn}{dx}\right)_{x = \mathrm{W}}$ with equations (16.3.3) and (16.3.5) as

$$\mathrm{I}_{\mathrm{diff}} = \mathrm{A} \, \eta \, q \, \mathrm{N}_0 \frac{\alpha \mathrm{L}_n}{1 + \alpha \mathrm{L}_n} \exp(-\alpha \mathrm{W}) ; \quad \cdots \quad (16.3.6)$$

Hence the photogenerated d.c. current under constant level of illuminations is obtained from (16.3.2) and (16.3.6) as

$$\mathrm{I}_{ph} = \mathrm{A} \, \mathrm{N}_0 \, q\eta \left[1 - \frac{\exp(-\alpha \mathrm{W})}{1 + \alpha \mathrm{L}_n}\right] \quad \cdots \quad (16.3.7)$$

Note that the total photocurrent is proportional to the incident photon density. For higher external quantum efficiency defined by $\eta_{ext} = (\mathrm{I}_{ph}/\mathrm{N}_0 \mathrm{A} q)$, it is desirable to have $\alpha \mathrm{W} \gg 1$ (higher absorption) and $\alpha \mathrm{L}_n \gg 1$ (higher generation compared to recombination). For longer wavelengths to satisfy $\alpha \mathrm{W} \gg 1$, one may require large W with the associated penalty of greater transit time delay.

To examine the transit time effect, we assume (for simplicity) that, the optical photon density is varying with time in the form $\mathrm{N}_0 \exp j\omega t$ and that, all the photons are absorbed at the depletion boundary at $x = 0$. This causes generation of $\mathrm{N}_0 \eta \, \alpha \exp(j\omega t)$ number of hole electron pairs per second on this surface, whereform one species travels with saturated velocity v_s through the intrinsic region. Assming no recombination, these carriers reach at x when a time (x/v_s) has elapsed. That is, the carriers that were generated on the $x=0$ plane at time $(t-x/v_s)$ reaches x at time t. So the conduction current density as a function of x is

$$J(x)=N_0 q\eta\,\alpha\,\frac{v_s}{W}\exp j\omega\,(t-x/v_s) \qquad \ldots \qquad (16.3.8)$$

Hence the total conduction current due to all carriers in the intrinsic region at time t is

$$J_c=\frac{N_0\, q\eta\,\alpha\, v_s}{W}\int_0^{w}\exp j\omega\,(t-x/v_s)\,dx$$

$$=N_0\, q\eta\,\alpha\, v_s\left[\frac{1-\exp j\omega t_r}{j\omega t_r}\right]\exp\,(j\omega t) \qquad \ldots \qquad (16.3.9)$$

where $t_r=(W/v_s)$, the transit time through the intrinsic region. The expression within the square brackets in equation (16.3.9) above has the form $[\sin\,\omega t_r/2]\,/\,(\omega t_r/2)$, which is maximum when $t_r=0$ and has the first zero when $wt_r=2\pi$ while the phase lags linearly as $(wt_r/2)$ radians. When $wt_r=2{\cdot}4$ radians the magnitude of this factor falls by $3db$ and the phase lags by about $(2\pi/5)$.

The total current is the sum of the conduction current and the displacement current while the displacement current depends on the junction voltage change. Hence ignoring the series resistance, the short circuit current density is given by equation (16.3.9). This expression shows that the transit time reduces the photoresponse and produces dispersion of the signal. Whereas $\alpha W \gg 1$ for more absorption and $\alpha L_n \gg 1$ for better collection is necessary. A reasonable trade off value of $W=1/\alpha$ is used. This gives for P-I-N diode a $3db$ bandwidth as

$$f_{3db}=\frac{2{\cdot}4}{2\pi\, t_r}=0{\cdot}4\;v_s/W=0{\cdot}4\;\alpha\, v_s \qquad \ldots \qquad (16.3.10)$$

P-N Junction Diode :

For a P-N junction photodiode on the other hand, the depletion region is thin. Assuming that there is no built-in field in the bulk to assist the collection of carriers, the total transit time for the minority carriers is mainly due to diffusion in the bulk. High intensity caused bulk generation would be highest on the illuminated end. If this is N-type, then the diffusion process determined cut-off frequency is given from equation (7.3.3) by

$$f_{\text{diff}}=\frac{\pi}{8}\,\frac{D_p}{W^2}=0{\cdot}4\alpha^2\, D_p \qquad \ldots \qquad (16.3.11)$$

where W is width of the *n*-type bulk on the illuminated end. $D_n > D_p$, therefore a P-N diode is better than an N-P diode in this respect (cf. fig. 16.3.2*b*).

Two extreme cases, the P-I-N diode, in which the carriers are collected mainly due to drift and the P-N diode in which the transit is mainly due to diffusion has been considered above. For the intermediate ones, in which there is doping gradient such as N^+-N-P-P^+ the cut-off frequency lies between these two limits.

That is, $0{\cdot}4_{\alpha}\, v_s > f_c > 0{\cdot}4\alpha^2\, D_p$ ··· (16.3.12)

This is the intrinsic cut-off frequency as determined by the carrier transport mechanism. This does not take account of the extrinsic elements such as the junction plus header capacitance and bulk plus contact series resistances. Normally it is these extrinsic elements that put the limit on the frequency performance.

Point Contact Diode :

In this respect, the point contact photodiodes have smallest capacitance. The active area [cf. fig. 16.3.2(c)] being a small volume around the contact point, it has lower drift time that gives high intrinsic cut-off frequency. Matched with its lower capacitance, it is suitable for very high modulation frequencies. But the degree of success in focusing the optical radiation upon the contact point determines the external quantum efficiency.

Metal-Semiconductor Diode :

A metal semiconductor photodiode is a reverse biased Schottky barrier diode. The optical signal is incident on the metal electrode of thickness less than 100Å, practically that eleminates absorption losses in the metal film. An antireflection coating [cf. fig. 16.3.2(d)] is always used to minimise large reflection losses from the metal surface. These diodes can be used in either electron emission mode or as a normal photodiode or as an avalanche photodiode depending on the excitation wavelength and applied reverse bias.

In electron emissiom mode, the photon energy should lie between ϕ_{bn} and E_g. So that the light incident on the metal causes electron emission over the barrier ϕ_{bn} to the semiconductor. The reverse bias voltage should be lower than the junction breakdown voltage. This causes metal emitted electrons to be collected by the semiconductor beyond the depletion region where these are majority carriers. This method of excitation is extensively used in the determination of ϕ_{bn} and in the study of hot electron transport through metal films.

To use a metal-semiconductor diode as a normal photodiode, the photon energy should be higher than the semiconductor band gap while the reverse bias is kept below the junction breakdown voltage. The holes and electrons are excited, mostly in the depletion region of the Schottky barriers. The carrier flow is assisted by the strong Field in the depletion region. Hence the diode external quantum efficiency and speed are comparable to P-I-N diodes.

To use the metal-semiconductor diode as an avalanche photodiode, the photon energy is above the band gap energy usually while the applied reverse bias is a little less than the junction breakdown voltage. This gives high current multiplication and improves the external quantum efficiency appreciably. Used with well-designed guard ring as shown in fig. 16.3 2(d), increased speed of response is obtained due to avalanche multiplication of carriers in the central depletion region.

Signal to Noise Ratio :

The photodiode equivalent circuit is shown in fig. (16.3.3). It consists of the photo-generated current generator I_p (ω) feeding the junction plus header capacitance C. The reverse biased a.c resistance being high is ignored while R is the series resistance due to the bulk and contact. The noise sources, discussed below, are also shown in this figure.

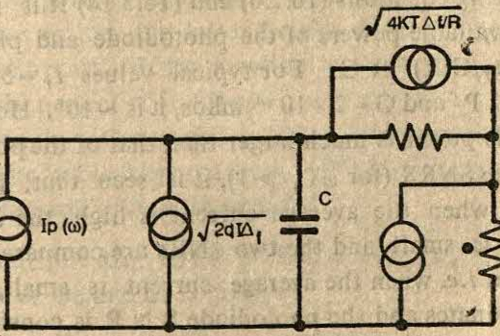


Fig. 16.3.3. Shows the photodiode equivalent circuit including the noise sources.

Assuming that the photo-generated carriers are swept out by the depletion region field before they can recombine, the photodiode signal current can be written as

$$I_{ph}(\omega) = A\eta q\, P_1 / \sqrt{2}\, hf \qquad \cdots \quad (16.3.13)$$

The signal current through the matched load R, assuming $2\omega CR \gg 1$ (which cannot normally be avoided), is $[I_p(\omega)/2\omega CR]$. So the available signal power is given by

$$P_a = [I_{ph}^2(\omega)/4\omega^2 C^2 R]$$

$$= \frac{R}{8}(A\eta q\, P_1/\omega\, C\, R\, hf)^2 \qquad \cdots \quad (16.3.14)$$

Recombination being ignored, only the generation originated shot noise is present. The mean square shot noise can be written in terms of d.c current $I = A\, q\eta .\, (P_s + P_L)/hf$ as $(2q\, I\, \Delta f)$, whose contribution to output noise power is $[R\, q\, I\, \Delta f/2(\omega CR)^2]$.

Assuming an equivalent temperature T_e to account for the load, the Johnson noise mean square current in matched load is $4KT_e\Delta f/R$. Ignoring the contact resistance noise the SNR for the diode can be written as

$$S/N = \left[\frac{A\eta m^2\, P_0}{4\,\Delta f\, hf}\right] \Big/ \left[1 + \frac{8\,KT_e}{q}\,\frac{(\omega CR)^2}{RI}\right] \qquad \cdots \quad (16.3.15)$$

Comparing equations (16.2.6) and (16.3.14) it is seen that, the ratio of the available powers of the photodiode and photoconductive detector is $(t_r/C\,R)^2\, R\, G$. For typical values $t_r = 50\ n$ sec, R= 1¯ohom, C= 2 Pf and $G = 2 \times 10^{-8}$ mhos, it is $\sim 10^6$. Hence the photodiode available power is much larger than that of the photoconductor. Comapring the SNRS (for $\omega T_0 \gg 1$), it is seen that, for high level detection *i.e.* when the average current is high, the Johnson noise contribution is small and the two SNRs are comparable. For low level detection *i.e.* when the average current is small, the Johnson noise predominates and the photodiode S N R is considerably higher than that of the photoconductor. For infrared detectors (the material used having low band gap), the diode reverse bias a.c resistance cannot be ignored and the equivalent resistance is to be considered particularly for low excitation levels.

The photodiodes have found ample use in mixing of optical signals at high intensity levels. Photoconductors however, have been extensively used in the far infrared and lower modulation frequencies because of the non-availability of a cheaper substitute.

16.4 Avalanche Photodiode

Avalanche photodiodes (APDs), are junction diodes operated under high reverse bias voltages, less than the junction breakdown voltage V_{br}. High bias field causes carrier multiplication due to collision ionisation of primary electrons and holes generated by photo-excitation. Substantial current gains are achieved even at MW frequencies and high external quantum efficiencies are obtained. Current gain bandwidth product as high as 100 GHz with maximum current gains of 150 has been reported (cf. ref. 1), at low intensity levels. The realisation of high gain bandwidth product is possible provided, the radiation generated primary carriers mostly flow through the high field region in the shortest possible time. So the collision ionisation process needs to be confined in the carrier flow path alone. This is achieved by using suitable constructional features as shown in fig. (16.4.1) ; such that, spatial uniformity of carrier multiplication in the light sensitive area is obtained. The guard ring used in the diodes in this figure ensures elimination of 'microplasmas' that may occur, in small areas at lower breakdown voltages. The guard ring having a lower impurity gradient and larger radius of curvature than the central active region, the latter breaks down earlier. This also reduces leakage currents. If the junction intersects the surfaces at an angle as in fig. (b) a mesa or in (d) contoured surface, lower fields are obtained at the edges. Contoured surface junction occurring inside the medium minimises leakage currents. These contours can be fabricated by out-diffusion of dopant material.

Current multiplication in the prebreakdown ($V_a < V_{br}$) region of P-N junctions of Ge and Si is given by

$$M_0 = 1 \Big/ \left[1 - \left(\frac{V_a}{V_{br}}\right)^m\right] \quad \cdots \quad (16.4.1)$$

where the value of m ranges for Si, between 1·5 to 4 and for Ge between 2·5 to 8. In reverse biased junctions there is leakage current as also unsuccessfully shielded radiation induced current. The total, in the absence of optical signal to be detected, is called dark current I_d. This also gets multiplied to I_{md} due to collision ionisation along with the optical signal generated carrier current I_{ph}. The multiplication to be measured is

$$M_{ph}=\frac{I-I_{md}}{I_{ph}}=\frac{1}{1-\left(\frac{V_a-IR}{V_{br}}\right)^m} \quad \cdots \quad (16.4.2)$$

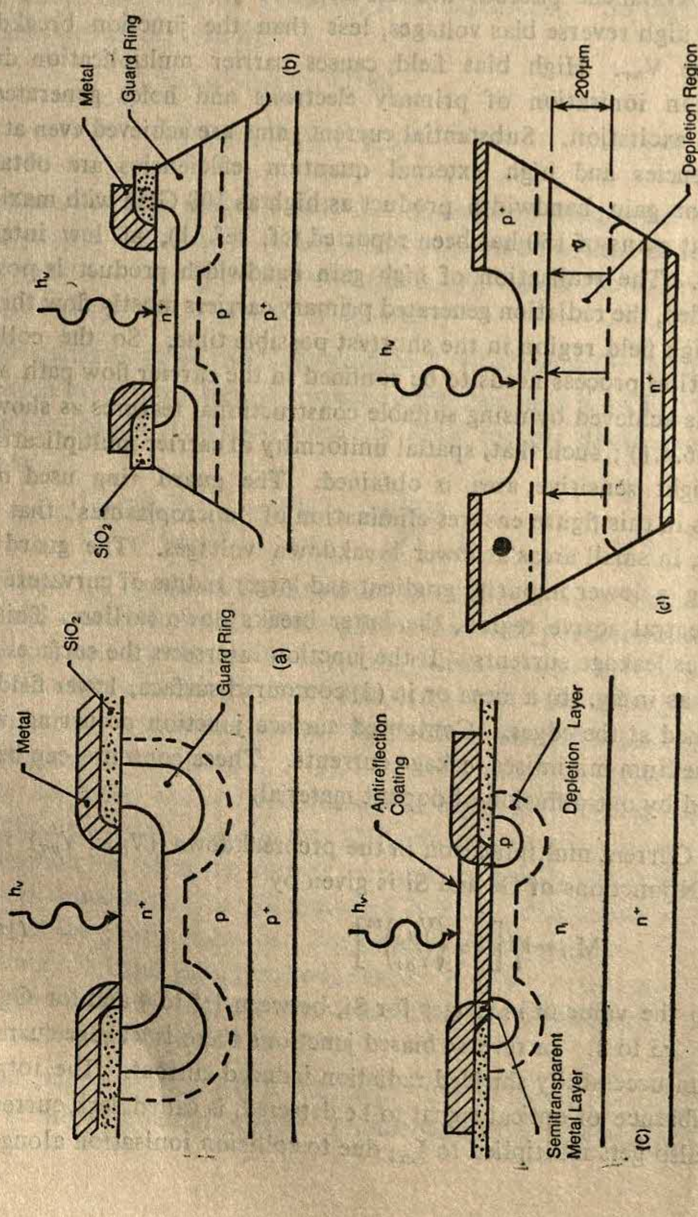


Fig. 16.4.1 (a—d). Construction of some avalanche photodiodes.

where I is the total output current in the event of multiplication, with the optical signal on, V_a the applied voltage and IR is the voltage drop in the dynamic resistance of the diode. The dynamic resistance arises from the carrier flow process, diode series resistance and load resistance. Equation (16.4.2) shows that, at high current levels the current gain saturates due to electric field lowering caused lowering of multiplication. For high signal levels $I \geqslant I_{md}$ and we have for V_a near V_{br} from equation (16.4.2)

$$M_{ph} = (I/I_{ph}) = (V_{br}/mIR) \quad \ldots \quad (16.4.3)$$

$$i.e. \quad I = \sqrt{(I_{ph}\, V_{br}/mR)} \quad \ldots \quad (16.4.4)$$

That is the multiplied maximum output current obtainable varies as the square root of the photo-generated current. For low signal levels on the other hand, $I_{ph} \leqslant I_d$, but $I = M_{ph}\,(I_d + I_{ph})$.

which with $M_{ph} = \dfrac{V_{br}}{mIR}$

gives $$I = \sqrt{(V_{br}\, I_d/mR)}\,(1 + I_{ph}/2\, I_d)$$

hence $$\frac{\delta I}{\delta I_{ph}} = \sqrt{(V_{br}/m\, RI_d)}/2 \quad \ldots \quad (16.4.5)$$

Equation (16.4.5) shows that the maximum multiplication is lower for higher dark current at low signal levels. Cooling yields lower dark currents hence higher multiplication factor that is particularly useful for diodes having high V_{br}.

Multiplication of carrier depends largely on the ionisation coefficients of holes and electrons, which are in turn field dependent (cf. ref. 2 and art. 12.1). In Ge ionisation coefficients for electrons are nearly equal to those of holes, in Si, and GaAs the ionisation coefficients for electrons are higher* than those of holes. These differences in ionisation coefficients greatly influence the gain and noise performance of the APDs. The spatial and temporal build-up of avalanche, due to one-carrier type being dominant, is different from both the carrier types taking part equally in the multiplication process. For example, in a nearly constant field region avalanche, where one carrier type is producing more than the other per cm, say the dominant one is the electron carriers. An electron starts at $t = 0$ and suffers a collision causing a pair generation a little ahead of $x = W$ (cf. fig. 16.3.1) on its transit through the depletion region (towards the left). The hole starts moving to the P-side while now there are two electrons. These two electrons in their turn produce two pairs of electrons and holes. This goes on until the electrons have reached $x = 0$ at time t_r, the electron transit time, when the induced short circuit current is maximum. The current thereafter falls until the last hole has reached the opposite junction edge at time $T = t_r + T_p$. The current gain obtainable in a situation like this is called unilateral gain which is seen to be independent of excitation frequency upto a 3 *db* band width of $(2{\cdot}8/\pi T)$. In this case the multiplication gain takes the form exp (αW) where α is the ionisation coefficient of the dominant carrier and is assumed here to be field independent. For steady state excitation a large number of electrons are present within the junction, so the statistical fluctuations in carrier generation is relatively weak. Low 'excess noise' is thus expected.

---

* S.M. Sze gives for GaAs the ionisation coefficients as equal but noise measurements indicate electrons have higher ionisation coefficients.

For both types of carrier taking part in collision ionisation equally on the other hand, gives a feed-back mechanism. Secondary holes generated near $x = 0$ move towards $x = W$ to generate a new electron near a location where the primary electron started the avalanche. Thus the current persists much longer than the transit times of holes plus that of electrons. For equal ionisation rates for holes and electrons the multiplication gain takes the form

$$M = [1/(1 - \alpha W)].$$

For αW tending to one. *i.e.* on the average one-carrier pair produced per passing carrier, gives very high gain. At any one instant then, only a small number of carriers are present in the high field region. The feedback mechanism responsible for high gain is then easily disturbed due to statistical fluctuations in the ionisation coefficient. Hence larger current fluctuations are expected that is higher excess noise in this case is obtained. It is therefore seen that when one-carrier is dominant in avalanche multiplication the noise is much less than when both the types of carriers are about equally active in pair generation.

For ideal APD to have low noise and large gain up to high modulation frequencies, one-ionising carrier type is necessary. The a.c. current multiplication gain is seen to be of the form (cf. ref. 1)

$M(W) = M_0/(1 + j\omega\, M_0\, T_{eff})$ where T_{eff} is an effective time constant proportional to the average value of the time constant T. Therefore for practical Ge and Si diodes the current gain bandwidth product is limited to $(1/2\pi T_{eff})$. It can be easily visualised that for short pulse duration (T_{pl}), the multiplied peak photo current $I = I_{ph} T_{pl}/T_{eff}$. Therefore high current gain bandwidth product requires narrow multiplication regions, that gives low T_{eff}. In such cases materials like GaAs and In Sb with one-type ionising carrier give high current gains and low noise but are not limited by gain bandwidth product. While low noise silicon APD requires wide multiplication region for high gain, since for one-carrier ionisation it requires comparatively low field operation. This causes gain bandwidth product limitation due to increased transit time incurred in obtaining high gain.

The equivalent circuit of the avalanche photodiode can be drawn as shown in fig. (16.4.2). In this figure $I_{ph}(\omega)$ represents the photo-generated current as a function of frequency which, ignoring the transit time effect can be expressed as in equation (16.3.13). The transit time effect being taken into account with current multiplication. The current amplifier M represents the avalanche multiplication of photodiode currents. This amplifier has frequency dependent characteristics. $\sqrt{\overline{i^2_n}}$ is the r m s shot noise current source. R is the dynamic resistance and R_L the load including the input of the amplifier that follows the detector.

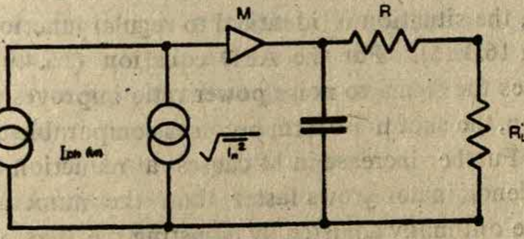


Fig. 16.4.2. Shows the equivalent circuit for APD with noise generator.

The signal generated available power in the load, therefore can be written as

$$P_a = \frac{I_{ph}{}^2(\omega)\, M^2\, R_L}{[1 + \omega^2\, C^2\, (R_S + R_L)^2]} \quad \cdots \quad (16.4.6)$$

The average power in the load due to shot noise can be written in terms of the d.c current as

$$P_{n1} = 2q\,I\,M^n\,R_L\,\Delta f/[1+\omega^2\,C^2\,(R_S+R_L)^2] \quad \cdots \quad (16.4.7)$$

where for two-carrier dominance in multiplication process $n=3$ while for one-carrier dominance in avalanche $2 < n < 4$.

The thermal noise power in the load can be written in terms of an equivalent temperature T_e as

$$P_{n2} = 4\,KT_e\,\Delta f/R_L \quad \cdots \quad (16.4.8)$$

Using equation (16.4.6) to (16.4.8) the signal to noise power ratio can be written as

$$(S/N) = \frac{I_{ph}^2}{2q\,I\,\Delta f}\left[\frac{M^2}{M^n + \frac{2kT_e}{R_L\,Iq}\left(1+\omega^2 C^2\;\overline{R_S+R_L}^{\,2}\right)}\right] \quad \cdots \quad (16.4.9)$$

When $M=1$, the situation is identical to regular junction photodiode (cf. equation 16.3.15). For the APD equation (16.4.9) shows that as M increases the signal to noise power ratio improves until, at low level detection, the shot noise term becomes comparable to the thermal noise term. Further increase in M causes a reduction in (S/N) ratio because the denominator grows faster than the numerator. If M is assumed to be optimally adjusted by adjusting the bias voltage, such that, the term due to shot noise equals that due to thermal noise in the denominator of (16.4.9), then the minimum detectable power obtained is a factor of M lower than that of the regular diodes. For Ge values of $M \sim 10$ at $I = 530$ micro amp has been seen to give highest signal to noise ratio of 20 db. At higher values of M the avalanche noise power increases faster than the output signal power in this material (cf. ref. 1).

Schottky barrier diodes of N-type silicon with guard ring has very low leakage current. These can be essentially operated as one-carrier diode hence are useful for high sensitivity and low noise with high gain bandwidth product. The small light sensitive junction areas of high speed regular photodiodes reguire precise focusing of the incoming signal. Silicon APD with high current gain and low excess noise are available for wavelength ranges from 0·4 to 1 micron. Ge APDs have higher excess noise and higher leakage current. If cooled these can be used upto 1·5 microns. When APDs are used for pulse detection, M should be optimised for highest acceptable pulse detection rates rather than for best S/N ratio. It is known that higher sensitivities can be reached in that case than the value S/N ratio approach gives (cf. ref. 1).

16.5 Heterojunction Diodes

In increasing the speed, sensitivity and the minimum detectable power in the near infrared to the near UV range (1·4 *e*V—3·1 *e*V), heterojunction detectors recently (cf. ref. 3) have proved to be quite promising. Using heterostructure diodes particularly using GaAs has shown possibilities of having a few orders of magnitude lower dark current, even without the necessity of cooling for applications in the near infrared range. GaAs has higher band gap hence has higher reverse bias withstanding capabilities. In heterostructure diodes control of collection of photo-generated carriers and elimination of surface recombination is possible. The high energy photons incident on the surface are either not allowed to excite carriers very near the surface by using wider band gap material at the surface, or the excited carriers are dragged inside faster than the carrier recombination velocity. This can be achieved with, a large built-in field using variable band gap material window, that cannot be obtained with doping gradient in Si or in Ge.

The heterostructure diodes that have attracted attention are shown in fig. (16.5.1). Absence or a favourable presence of discontinuity in the P-N or N-P heterojunction energy bands ensures complete separation [cf. fig. 16.5.1(a)] of non-equilibrium carriers generated both in the wide gap and in the narrow gap material of the pair. This junction is usable in the photon energy range $E_{g1} < hf < E_{g2}$, such that, the photons having energies lower than E_{g2}, penetrate through the wide band gap window and excite electron hole pairs, well inside the junction region. Thus practically no recombination loss of excited carriers occurs on the illuminated surface. The total drift plus diffusion contributed minority current can be given approximately by equation (16.3.7) which indicates that, for efficient collection $\alpha L_p \gg 1$ is required. That is, light doping, which is expected to give less tampered hence comparatively larger L_p, is to be used for the narrow gap material. N-type GaAs substrate

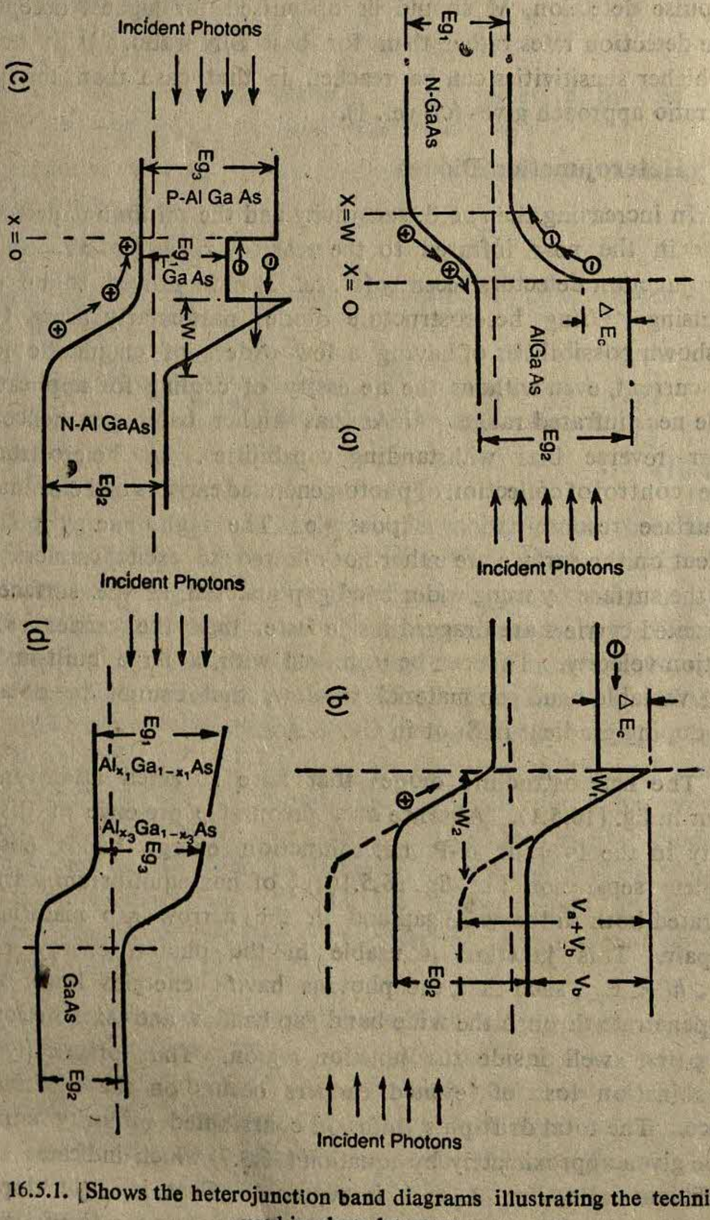


Fig. 16.5.1. [Shows the heterojunction band diagrams illustrating the techniques used in photodetectors.

with suitable Al content for P-type $Al_xGa_{1-x}As$ is to be used to obtain high spectral response over larger band. Experimental results (ref. 3) show that only a lightly doped ($N_a \sim 10^{14}$/cc) narrow gap region gives high minority carrier collection ($\eta_{ext} = 0{\cdot}95$) in the spectral region of 0·65 to 0·89 microns (1·4—1·9 ev). The thickness and doping of the N-GaAs should be proper (0·5—1·0 micron), so that the optical absorption beyond the space charge layer at distances greater than a few diffusion lengths is avoided.

As another example, the control of minority current through the conduction band discontinuity may be mentioned [cf. fig. 16.5.1(b)]. In this case the photon energy also lies in the range $E_{g1} < hf < E_{g2}$. The current through the barrier is largely tunnel current. As expected the photocurrent obtained shows a dependence of the form

$$I_{ph} = I_{op} \exp\,[-\beta(V_b \pm V_a)^{-\frac{1}{2}}] \qquad \cdots \quad (16.5.1)$$

where β is as given in equation (15.2.9). Equation (16.5.1) shows that the barrier $\triangle E_c$ transparency is controllable by applied reverse bias. Experiment (cf. ref. 3) shows that the external quantum efficiency increases with applied reverse bias upto a certain extent. In this case low doping of the wide gap N-type material is seen to give best results. This example shows though a great majority of heterojunctions have non-zero band discontinuity suitable doping of the materials of certain junctions can give broad band spectral sensitive photodetectors. The bias dependent carrier separation built-in mechanism, enables bias dependent spectral sensitivity control over the whole of its spectral range.

For increasing the speed of response and sensitivity the double heterostructure diodes also hold some promise. As shown in fig. 16.5.1(c) it consists of a P-type GaAs sandwiched between a P-type $Al_xGa_{1-x}As$ and an N-type $Al_xGa_{1-x}As$. Another diode [cf. 16.5.1(d)] consists of a P-type $Al_x{}^1Ga_{1-x}{}^1As$ followed by a P-type $Al_{x3}Ga_{1-x3}As$ forming the junction with N-type $Al_{x2}Ga_{1-x2}As$ where $x_1 > x_3 > x_2$. Increase of molefraction x of aluminium in GaAs lattice increases the band gap of GaAs from 1·45 to 1·9 eV at $x = 0{\cdot}4$. For higher x the material becomes indirect gap having $E_g = 2{\cdot}2$ eV at $x = 1$ eV. In the first one the wide band gap on the front surface allows the radiation, a window, that enters to get absorbed in GaAs. The barrier at position $x = 0$ [cf. fig. 16.5.1(c)] causes the electrons to get reflected into GaAs allowing their unidirectional flow to the N-type $Al_xGa_{1-x}As$. This occurs to the extent the barrier becomes transparent by applied reverse bias. The holes do not find any barrier hence they move to the P-side. Photon energy in this case may range from $E_{g1} < hf < E_{g3}$. This eleminates surface recombination and causes separation giving controllable collection. Experiment shows (cf. ref. 3) that light absorbed in the pit region gives largest photocurrent when the pit width is ~ 20 micron. In this case long λ limit is seen to correspond to 1·35 *e*V. The second structure (fig. *d*) having a graded band gap offers a window having a built-in field for the electron carriers different from that for the holes. These fields are determined by the slopes of the respective band edges. If the mole fraction '*x*' of Al changes from 0·4 to near zero in a distance of typically 0·3 micron, then a field greater than 10^4 V/cm is obtained. This causes the high energy photon-generated N-type carriers at the front surface to move out faster than their surface recombination velocity. This improves the shortwave generated carrier collection. On the other hand carrier generation due to long λ photons (corresponding to 1·45 *e*V), are absorbed well inside GaAs where the mole fraction $x_2 = 0$. In another version on the front surface a thin protective layer of $Al_xGa_{1-x}As$ with $x = 0{\cdot}9$, that is transparent upto 3·1 *e*V (being indirect gap), has been used that extended the spectral range covering 0·4 to 0·9 micron. Thus we see that the heterostructure diodes using GaAs have encouraging possibilities, particularly the leakage current being orders of magnitude small ; electron mobility high, high breakdown that allows large bias application including isolation possibility on GaAs substrate, promises for high speed, high gain bandwidth product and high sensitivity. Low noise my also result since higher electron ionisation coefficient is seen on noise measurements in GaAs diodes.

References

1. Laser Handbook Vol. 1 by F.T. ARECCHI & E.O. SCHUL-DUBOIS. North-Holland Publishing Company 1972.
2. Physics of Semiconductor Devices by S.M. Sze, p. 60. John Wiley & Sons Inc. 1969.
3. Semiconductor Optoelectronics, Marian A Herman. John Wiley & Sons (1980).

INDEX

[A]

- Abrupt junction 66, 417
  - anisotype 410
  - isotype 417
- Absorption coefficient 131, 435
  - Cd Te, Ga As, Ga P, Ge, Si
- Absorption limit 63
- Acceptor 48
- Acceleration 21
- Accumulation 137, 256, 352
- Acoustic phonon 43
- Active region 212
- A C Impedance diode 110
- Admittance junction diode 110
  - transistor 190
  - FET 239
  - Read diode 323
- Al-Ga-As 454
- Allowed 16
- Alpha variation 165
  - cut-off frequency 183
- Aluminum 249, 253
- Ambipolar mobility 436
- Anti-symmetric wave function 32
- Application of Fermi-statistics 34
- Avalanche breakdown 122, 173
  - multiplication 123, 312, 334
  - photo-diode 447
  - region, equations 314, 316
- Arsenic 47

[B]

- Band theory 13
  - conduction, valence 12
  - diagram 30, 137
  - gap 31, 56, 421
  - overlap 29
- Barrier height 145
  - injection 342
- Base charge 207, 214
  - diffusion 252
  - excess charge 208
  - recombination 105, 119, 177
  - resistance 178
  - stored charge 112
  - transit time 112
  - transport factor 104
  - width modulation 179
- Beta cut-off frequency 196
  - measurement 198
- Bipolar transistor a.c operation 177
  - technology 254
- Body centered 3
- Boltzmann 37
- Bose-Einstein Statistics 32
- Bragg reflection 20, 24
- Break down junction 123
  - curvature effect 124
  - temperature effect 124
- Boundary conditions 19
  - concentrations 74
- Bravais lattice 3
- Brillouin zone 18, 30
- Bucher's Equal-area rule 359
- Buffered HF 248
- Built-in field 68, 73, 86
  - voltage determination 82
- Bulk negative diff. conductivity 348
- Burried layer 251

[C]

- Carrier decay 76
- Capacitance
  - junctions 69, 70, 412, 419
  - P-I-N diode 126
  - avalanche region 323
  - measurement 200
- Capture rate 64
- Carrier decay process 77
  - transport 104
  - velocity 43
- CdS, CdSe 435
- Circuit (MW) IMPATT 338
  - TRAPATT 335, 337
  - Gunn 381, 382
- Channel current 229, 259
  - thickness 225

- Charge & potential
  - in MOS 266
  - in P N junction 67, 412
- Charge control equns 209
- Charge relaxation 78
  - storage 112
  - storage diode 121
- CMOS 288, 296
- Collection efficiency 163
- Collition ionisation 122
- Common base 162
  - emitter parameters 188
- Composition of V_T 270
- Complementary Error function 244
- Contoured surface 447
- Covalent crystal 47
- Crystal momentum 22
- Concentration extrinsic 54
  - intrinsic 47
  - minority 75
- Conductance (IMPATT) 324, 328
- Conduction band
  - effective density of states 37
- Conductivity 48
  - effective mass 31
  - surface 256
- Contact potential 68
- Continuity Equations 80, 424
  - solutions 101
- Current density expressions 58
  - graded heterostructure 424
- Current gain 162, 436
  - band width product 198, 447, 451
- Current limiter diode 235
- Current pulse (Read diode) 313
- Current transient (Junction diode) 113
- Current voltage relations 99, 144, 147, 230, 260, 413, 419
- Current voltage characteristics
  - Bipolar transistor 166
  - Junction FET 231
  - MOS FET 260
- Cut off frequency (IMPATT) 327
  - Junction transis or 185, 196
    - FET 241
  - Photo diode 443
- Cut-off region
  - switching transistor 212
  - wave length 439

[D]

- Dark current 431
- Deformation potential 39
- Deep impurities 434
  - levels 156
- Deep level trap 157
- Degenerate 37
- Delay time 206
- Delayed mode 371
- D-MOS 308
- Density of states 37
  - effective mass 62
- Depletion region 67, 256
  - capacitance 81
  - deep 276, 279
  - mode MOST 263
  - width 69
- Detectors (photo) 428
- Detectivity 429
- Die-bonding 253
- Diffusion 56, 243
  - Capacitance 116
  - length 105
  - limited diode 105
  - model 145, 415
  - profiles 245
- Dimond structure 7
- Diode fast recovery 122
  - met-sem 444
  - point contact 149, 444
  - P-I-N 125
  - Schottky 142
  - S-C-L 150
  - step recovery 122
  - varactor 152
  - zener 124
- Diode equation 100, 147, 415, 419
- Direct band gap 30, 347, 421
- Domain 351
  - shape 363
  - voltage 364, 403
- Doping profile 81

Double diffused MOS 308
Drain current 260
Drift mobility 44
Drift region 312, 319
  in BJT 98
  capacitance 323
  velocity 43, 86
Dynode 430

[E]

Early effect 179
Electron beam heating 250
Ebers Moll model 207
Effective mass 23, 62
  conductivity 31
  longitudinal & transverse 30,62
Effect of back gate bias 282
  temperature 46, 283
  on breakdown 125
  surface states on MOS 281
  MOST current 283
  unswept layer 332
Efficiency external 445
  Gunn diode 372, 376
  Read diode 328
  quantum 431
Electron affinity 133
  number of 36
Elementary theory (met-sem) 133
Emission model 143
  recombination 415
Emitter diffusion 252
Entrancement mode 262
Epitaxy 246
Epitaxial reactor 247
Equivalent circuit, diode 127, 153
  Jn. transistor 191, 200
  Jn. FET 241
  Gunn diode 369
  IMPATT 339-340
  MOS Capacitor 278
  MOSFET 404
  Photo detector 438, 445
  APD 451
Equal area rule 359
Error function (complementary) 244
Etch rate 63
Extrinsic parameters (CB) 187
  Common emitter 196
External cut-off frequency 441
Extrinsic semiconductor 40
  Fermi level in 52
  photoconductor 434

[F]

Face centered 3
Fall time 206
Fast recovery diode 121
Fast surface states 280
FED 236
Fermi function 32, 52
  IMREF (quasi) 59, 64
Field 44, 315, 342
  built in 88, 112
  changes 279
  depletion region 226
  distribution 350
  excess 334
  effect diode 235
  quasi 423
  strength 257
  variation 88, 126
Figure of merit 234, 261
Flat-band voltage 275
Forbidden range 16
Forward bias cut-off 149
Fractional reclaimable charge 118
Frequency extrinsic 187
  intrinsic 185
  modulation 339
  optimum 327
Frequency of oscillation 202

[G]

Ga As absorption co-efficient 435
  crystal structure 7
  energy band drg. 30
    band gap 61, 421
  temp. variation 62
  effective mass 62
  heterojunction 454
  ionisation co-efficient 413
  mobility 349, 394
  MESFET 394

  - phonon-spectra 42
  - velocity field 44, 348
- Gain band width product 261
  - current 162
  - power 160, 192
- Ga P 421, 435
- Gate 228, 257
- Gauss' law 273
- Ge absorption co-efficient 435
  - crystal structure 7
  - energy band diagram 30
  - lattice constant 413
  - mobility 46, 63
  - phonon spectra 42
  - velocity field 44
- General expression for current 103
- Generalised P-N diode 88
- Generation recombination 79, 131, 435
- Graded heterostructure 421
- Gradual channel approx. 229
- Ground state 47
- Grown junction 66
- Growth rate 327, 352
- Guard ring 447
- Gunn device 347
  - modes 370-381
  - mode chart 383
  - mounting 381
  - oscillations 347

[H]

- Hall co-efficient 50
  - effect 49
  - mobility 51
- Haynes Shockley expt. 80, 130
- Heavy doping 76, 152
- Heterojunction diode 410, 453
  - current expressions 414, 417, 419, 424
  - continuity equation 424
  - graded structure 421
  - energy band model 411, 414, 418
  - effective diffusion length 425
  - interface states 413, 420, 424
  - minority injection 424
  - photo-diodes 453
- High field 45, 347
- High vacuum system 249
- Hole capture 65
- Hot electron 44
- Hyperabrupt junction 152

[I]

- IC's 248, 251
- Ideal space charge, field 67, 89
  - capacitance 279
- Image force 139
- Impact ionisation 122, 172, 312
  - ionisation rate 315, 449
- IMPATT diode 312
  - diff. structures 327
  - measuring circuit 338
  - mounting 338, 340
  - small signal admittance 323
  - large signal 330
- Imperfect cathode 356
- Implantation 303
  - ion distribution 304
- Impurity states 49
- Improved theory (met-sem) 139
- Imref 59
- Incremental neg-conductance 327
- Indirect band gap 421
- Infrared 428
- Inhibit operation 387
- Injection 74, 424
  - efficiency 100, 108, 417
- Interface states 413
  - charges 411
- Intervalley 46, 347
- Intravalley 46
- Intrinsic concentration 48
  - Fermi potential 267
  - photo-conductor 434
- Inversion 256, 269
- Ion core 46
- Ion implantation 303, 305
- Ionised 48, 54
- Ionisation co-efficient 449
- Ionisation Energy 49
- Isolation diffusion 251

[J]

Johnson noise 432
Junction anisotype 410
  abrupt 66, 410, 417
  graded 66
  FET 228
  hyperabrupt 152
  isotype 417
  metal-sem 133, 140
  transistor 158, 177

[K]

K power gain 178, 182, 195
KPR 248
Kronig-Penney model 14

[L]

Large signal IMPATT 330
Lattice match Ge-Ga As 413
Lattice points 1
  vibration 39
Leakage current 125
Life time electron 80
  hole 79
LSA mode 379
Linearly graded Jn. 66
Load driver pairs 287
Longitudinal mode 42
Lorentz force 50

[M]

Magnetoresistance effect 50
Majority carrier 48, 256
Maximum field 70, 123
  oscillation freq. 202
  power gain 162, 192, 195
  Q_{max} 154
  width (depletion) 269
Maxwellian distribution 144
Mean free path 40
  time 42
Measurement
  built-in-voltage 82
  doping profile 81
  Hall voltage 49
  Junction Cap 200
  Punch through voltage 170
  switching parameter 213
  transistor parameter 198
  transistor breakdown 174
MESFET 393
  equivalent circuit 404
  design considerations 406
Metal okide semiconductor
  structure 253
  transistor 258
Metal semiconductor 133, 139, 444
Metal work function 133
Metallisation 248, 253
Microwave circuit 312
  switch 127
Miller indices 4
Minimum 151, 327, 429
Minority carrier 48
  life time 80
  injection 73
  distribution 96
Mobility drift 44
  Hall 51
  temp. variation 63
Mode chart (gunn diode) 382
Modulation index 428
Monochromatic 429
Monolithic circuit 250
MOS capacitance 275, 279
  Technology 284, 300
MOST 258
Mott barrier 156
Mott Gurney law 151
Mounting IMPATT 339
  Gunn 381
Multiplication 172, 431, 447

[N]

N-channel depletion 263
  enhancement 260
Negative resistance 312, 342, 347
Neutrality 77
Nitride on silicon 302
Noise 340, 383, 432
  equivalent power 429
  figure 383
  (1/f) type 341
  (S/N) ratio
    341, 432, 438, 446, 452

Non-thermal 415
Non degenerate 37
N-P-N transistor 163
np product 46

[O]

Ohmic contact 76, 253
Optical mode 41, 42
  phonon 147, 434
Optical signal 428
Optimum frequency 327
Oscillation frequency 327
Out-diffusion 447
Out-put characteristics 166, 260, 263
Oxidation 246
Oxide thickness 284, 301

[P]

Pauli exclusion principle 26, 32
PDLT, PDLS 294
PELS, PELT 288, 291
Phasor relation (Read diode) 325
Phonon scattering 40
  spectra 42
Phosphorous 47
Photoconductor 434
  current gain 436
Photo detector 428
Photo-emissive detector 429
  multiplication 447
  multiplier 430
Photo-resist 248
Pinch-off 228
P-I-N diode 125
  photo-diode 441
P-N Junction 66, 86
  photo-diode 443
Point-contact diode 149
  photo-diode 444
Poisson's equation 160, 192
Polysilicon gate process 302
Power frequency limitation 334
Primitive vectors 1
Projected Range 304
Punch through 168

[Q]

Quality factor (Q) 154, 327
Quasi continuous modes 21
  Fermi level 64, 109
Quantum efficiency 132, 431, 435
  gain 436
  limit 433
  reflection 147
Quenched mode 373

[R]

Random 42
Read diode 312, 318
  admittance 323, 326
  efficiency 328
Reciprocal lattice 29
Recombination centers 122
Recombination generation 79, 437
  rate 65
  velocity 131
Rectifying contact 135, 138
Resonant loading 370
Reverse bias capacitance 126, 200
Richardson constant 144, 432
Rise time 219, 298
RWH mechanism 347

[S]

Satellite valley 347
Saturation region 212, 233
  JFET 225, 233
  MOSFET 260, 288
Schottky barrier diode 142
  effect 139
  emission 142
  emission diffusion 146
  gate FET 395
  phote-diode 444
Schrodinger equation 60
Schockley Read Model 85
Semiconductor degenerate 37
Sensitivity 429
Shallow level 434
Shot noise 432, 452
Silicon absorption co-efficient 435
  band diagram 30
  effective mass 62

  intrinsic conc. 48
  ionisation rate 315
  mobility 63
  phonon spectra 42
Silicon dioxide 246, 273
  nitride 302
Single trap level 84
Small signal alpha 185
  admittance (diode) 119
  transistor 190
  IMPATT 323
Snap back diode 122
SOS 305
Space charge growth 353, 378
  decay 379
Space-charge-limited 150
Storage phase 120
  delay 206
Stored & reclaimable charges 112
Surface potential and charges 269
Surface recombination velocity 131
Surface states in MOS 281
Switching characteristics diode 119
  transistor 206

[T]

Technology 243, 284, 300
TED signal processing 384
Temperature co-efficient 124
  effect 45, 61, 63, 234, 283
Tetragonal 2
Time evolution 352
Thermal current 415
Thermionic emission 143, 147
  emission diffusion 146
Transconductance 231, 260
Transfer characteristics 233, 262, 290, 293
Transient performance 299, 349
  response 205
Transient current 113
Transparent 139
Transistor Junction 158
  JFET 225
  MOSFET 255
  MESFET 393
Transit angle 322
Transit time 112, 322, 353
Transition region 67
Transport factor 104
Trap electron, level 64
  hole, 65
Tunnel 125, 139
Tunnelling model 147, 415
Turn off, on 206

[U]

Ultraviolet 428
Unilateral gains 194
Unit cell 2

[V]

Vacuum system 249
Valence band 27
  effective density of states 62
Varactor diode 152
Variation of alpha 165
Variation of IMREF 109
Velocity field characteristics 44
Velocity relaxation time 43
Visible region 428
Video detection 428

[W]

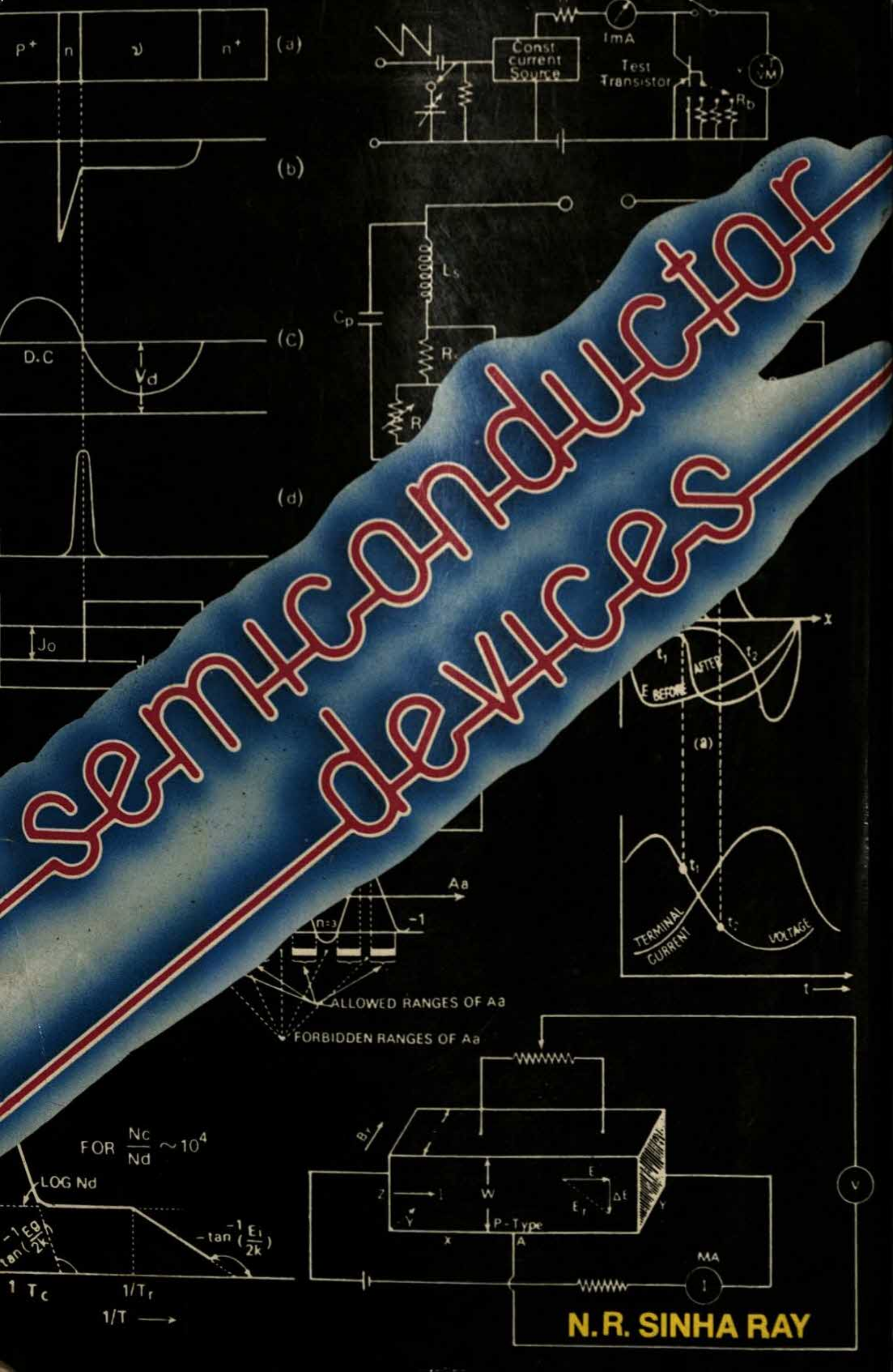
Wave equation 60
Wave vector 28
  factor 19
Work function 133

[X]

X'ray diffraction 20

[Z]

Zinc blende structure 7
Zone (Brillouin) 18



semiconductor devices

N. R. SINHA RAY